基于基因表达谱的肿瘤预测模型研究

李 辉[1], 王金莲[2]
(1. 北京工业大学计算机学院, 北京 100124; 2. 北京工业大学电控学院, 北京 100124)

摘 要: 本文从肿瘤基因表达谱分析入手, 研究并选取胃癌相关标志基因集合, 以此集合为基础抽取甄别肿瘤与正常组织的基因分类规则集, 进而建立起肿瘤预测模型. 首先, 以支持向量机为分类器用特征基因集合的样本识别率为适应度函数, 采用遗传算法对特征基因进行筛选. 然后用决策树抽取特征基因的规则集, 结合肿瘤分子生物学文献和生物实验建立肿瘤预测模型. 最后通过对胃癌基因表达谱数据的分析, 建立了胃癌预测模型, 结果表明该模型对胃癌分子生物学实验和临床诊断具有一定的指导意义和参考价值.

关键词: 胃癌; 遗传算法; 决策树; 支持向量机
中图分类号: TP18, TP391, Q617 文献标识码: A 文章编号: 0372-2112 (2008) 05-0989-04

Study of Tumor Molecular Prediction Model Based on Gene Expression Profiles

LI Hui[1], WANG Jin-lian[2]
(1. College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China;
2. College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China)

Abstract: Gene expression profiles of gastric cancer and counterpart normal tissues were analyzed with bioinformatics and machine learning methods to address the problem of discovery of biomarker genes and model of the tumor molecular diagnosis. Firstly, a support vector machine (SVM) was employed to find the feature gene subset with best classification performance for distinguishing cancerous tissues and the counterparts. And then, using genetic algorithm filter feature genes based on the classification rate of SVM. The decision tree was employed to extract rule subsets from the feature genes. These rules were tested by the test dataset with cross validation method, and the diagnosis model was constructed with these rules. The results indicate the gastric biomarkers and the diagnosis model would give more instruction in biological experiments and clinical diagnosis reference model.

Key words: gastric cancer; genetic algorithm; decision tree; support vector machine

1 引言

基因表达谱已广泛应用于肿瘤临床生物学行为预测, 如转移潜能[1]、药物敏感性[2]、预后判断[3]、肿瘤分子分型[4]、分子诊断标志和药物作用靶点等等. 1999 年 MIT 的 Golub[5] 博士利用基因表达谱, 研究了急性白血病不同亚型的自动识别问题. Hippo 等首次利用 Oligo 基因芯片分析了进展期胃癌组织与配对的癌旁正常组织的差异基因表达谱[6]. Boussioutas[7] 等人用 9381 个基因的表达谱比较了肠型胃癌、弥漫型胃癌、混合型胃癌以及癌旁的慢性胃炎、肠上皮化生组织的特异表达基因.

本文针对胃癌基因表达谱, 结合机器学习方法对胃癌分子标志以及预测模型进行研究. 本文分析的重点在于(1) 发现胃癌基因表达谱中和胃癌相关的标志基因; (2) 考察这些基因是否能作为区分肿瘤和正常组织的分子依据; (3) 挖掘这些基因之间的逻辑关系以及由这些关系组成的规则集是否可以区分肿瘤和正常组织; (4) 考察规则集中的基因表达模式和基因表达的阈值. 本文实验结果表明: 确存在区别于正常和肿瘤组织的异常表达的特征基因, 这些基因的表达水平可以对肿瘤与正常组织准确分类.

2 问题描述

由于基因表达谱数据的一个显著特点是样本少、维数高、数据不完整, 样本属性之间存在着大量复杂的相互关系, 因此找出影响样本表型信息的基因, 即样本分类特征基因就成为肿瘤基因表达谱分析的关键问题之一. 关键问题之二是提取特征基因规则集. 规则集是特征基因之间隐含的相互作用关系的描述, 并作为甄别肿瘤和正常组织的依据. 用决策树提取分类规则集法有如下优点: (1) 速度快、计算量小, 且易转化为分类规则. (2) 挖掘出的分类规则准确性高, 便于理解.

针对以上问题, 本文首先以 SVM 为分类器用特征基因集合的样本识别率为适应度函数, 用遗传算法对特征基因集合优化; 然后通过决策树训练分类特征基因, 抽取出蕴含于基因之间的逻辑关系和相互作用关系规则集, 最后结合生物学知识并以这些特征基因包含的规则集作为肿瘤预测模型. 具体实验原理见图 1.

收稿日期: 2007-05-08; 修回日期: 2008-01-13

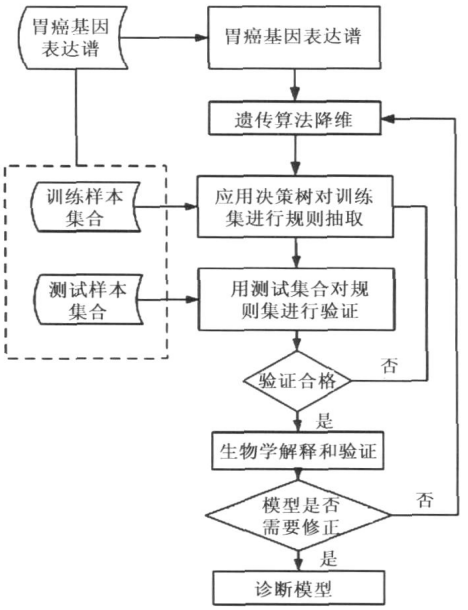

图 1 胃癌预测模型实验原理

3 实验数据

本文采用肠型胃癌差异基因表达谱数据, III, IV 期肠型胃癌患者 20 例与其配对的癌旁形态学正常组织 20 例, 21329 个基因的表达数据经处理后得到 1519 个大于 22fold 的差异表达基因.

4 特征基因的选取

虽然从基因组合的角度出发进行分类特征基因的选取更加符合生物学实际, 但是 1519 个基因构成的基因组合子集有 2^{1519}, 在如此巨大的特征子集空间进行最优搜索是非常困难的, 因此需要采用智能优化算法对特征空间进行搜索以找到最佳子集.

1519 个基因的解空间为 2^{1519}, 顺序浮动法[8] 是采用一种优化的部分集合来找出最优解的算法, 由于本文中的空间集合大, 很难分割出一种最佳集合, 如果集合选取过小, 则容易丢失最优解, 如果选取过大则空间复杂性就会增高. 信噪比[5] 对缺失值较敏感. 而且顺序浮动法和信噪比都是在指定的空间对每个解空间遍历, 其解空间的搜索速度就等于解空间的时间复杂度. 而遗传算法在整个集合中依据启发知识对全局寻优, 再加上合适的适应度函数以及选择、交叉、变异等算子选取都能提高搜索速度, 因此降低了时间复杂性. 由于遗传算法采用了编码操作, 它直接处理参数的编码集而不是解空间参数本身, 搜索过程不受空间复杂度的约束, 通过编码以及具有启发性知识的适应度函数进行空间搜索以寻求最优解, 相比顺序浮动法遗传算法的空间复杂度较低. 因此综合以上分析我们采用遗传算法对特征基因解空间进行搜索从而找到最佳分类基因集合.

遗传算法由四个要素组成: (1) 初始化种群; (2) 适应度函数选择; (3) 遗传算子(选择、交叉、变异) ; (4) 参数. 分别描述如下: (1) 初始化种群采用随机方法确定染色体的长度. 首先采用二进制编码技术将每个基因按照基因序号进行编码, 得到二进制子串, 将这些子串连成一个完整的染色体, 一条染色体代表一个特征基因集合. 其长度为基因集合个数, 1519 个基因需要用 11 位 2 进制数来表示($\lg_2^{1519}$ = 101568) , 每个基因的取值范围为 1~ $2^{10.568}$. 这样每条染色体编码就映射成了二进制串. 为了节省存储空间, 基因之间采用$ 符号来分割, 这样每个基因号就不用定长的二进制来表示了. 例如 10111 $ 101$ 10$ 0 分别表示第 23、5、2、0 号基因. (2) 适应度函数. 适应度函数是遗传算法的评价函数. 本文采用对残缺值不敏感的 SVM 算法作为适应度函数. SVM 由 Vapnik 等人基于统计学习理论采用结构风险最小化原理提出的机器学习算法[9] . SVM 把模式向量映射到高维特征空间而构造出一个最优分类超平面, 分类结果是基于各个样本值散落在超平面附近的距离来决策的, 因此样本的个别数据残缺并不影响整体的分类效果.

本文以 SVM 对特征基因集合的样本分类正确率作为遗传算法的适应度函数. 将样本集划分为训练集和测试集, 对测试集中的样本逐个进行类型识别, 并记录其分类错误率, 错误率低于 10% 作为选择基因的标准. 本文的 SVM 的核函数采用高斯核函数, 惩罚因子 C= 300.

SVM 评价基因集合的分类能力步骤如下:

第一步: 将胃癌及其对应的正常组织样本按近似 2 Bl 的比例分配在训练集和测试集中, 如图 2 所示.

图 2 样本划分

第二步: 训练集中的样本用作特征选取的学习样本, 经过学习得到决定样本类别特征的分类特征基因集合.

第三步: 用训练好的分类器作对测试集中的样本采用/ 留一交叉检验法0 (Leave2One2Out Cross Validation, LOOCV) 对样本类别预测. 重复该过程, 直到训练集上所有样本均有一次机会被测试为止. 记录所有被错分的样本数. (3) 遗传算子. 选择: 根据种群中个体的适应值, 采用赌轮选择. 交叉: 随机配对, 然后随机选择某一基因位置进行交叉. 变异: 随机选择某一个个体基因然后在 02 11111111111 范围内随机变异. 终止条件判断: 如果某一染色体的适应度达到最佳则遗传算法停止, 并输出结果. (4) 参数. 遗传算法中控制参数的选择非常重要, 一般控制参数包括初始种群大小 N, 交叉率 C_R, 变异率

P_R. N 的选择直接影响遗传算法能否找到最优解. 因为 N 太小会导致计算量增加, 收敛时间变长, 一般 N 取: 20 - 100. 交叉概率 C_R 控制着交叉操作的频率, C_R 太大, 会使高适应的种群遗失, C_R 太小搜索就会停止不前. 一般 C_R 取 014- 0199. 变异概率 P_R 是增大种群多样性的第二个因素, P_R 直接影响到算法的收敛性. P_R 增大会相应增加染色体的多样性和可选择性, 但 P_R 太大会使算法很难收敛. P_R 太小则难以产生新的染色体结构. 本文初始选 N= 20, C_R= 80%, P_R= 10%, 经实验发现搜索到的值并不理想, 为了找到最优解, 根据经验把交叉概率 C_R 调整到 97%, 把 PR 提高到 25%. 分类准确率有了较大提高. 利用遗传算法得到最佳分类基因子集后, 再用决策树对这些子集的样本识别能力进行测试, 最后选取识别率最高的那组参数作为最终的控制参数. 参数选择和决策树的识别率见表 1. 遗传算法见算法 1.

表 1　控制参数和决策树样本识别率对应表

种群数量	交叉率	变异率	决策树样本识别率
20	80%	17%	70%
20	70%	20%	70%
20	100%	25%	100

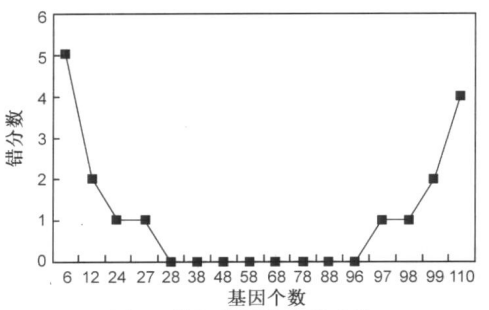

图 3　特征基因集合错分数

算法 1:

(1) generate an initial population F(o):= {g_x}(x= 1, 2, 3, , , A) which contained the gene randomly selected from the raw set F. (初始化种群从原始基因集合中随机选取 A 个基因作为初始集合 F(o))

(2) randomly select a pair of genes (i, j) from F(o), compute the union set of the pair of genes U(i, j); (从 F(o)中任意选择 2 个基因, 计算它们的并集 U(i, j))

(3) randomly select two genes (m, n) from F(oc) to produce offspring where g_m< U(i, j) & g_n< U(i, j), (从 F(oc)中随机选取 2 个小于并集值的基因作为子代成员)

(4) Randomly set a gene from F(o) to mutate as the offspring. (从 F(o)中随机确定一个基因进行变异)

(5) Compute the classification rate R(t) of F(o) randomly selecting from F(t)= F(o)+ F(oc), discard two maximum R(t). (从 F(t)= F(o)+ F(oc)中每次随机抽取 A 个基因并用 SVM 计算 A 个基因的分类正确率, 扔掉 SVM 分类正确率最小的 2 个基因)

(6) Return to step 2 (返回步骤 2)

(7) Compute R(t)= 90%, stop. (当 R(t)= 90%时, 算法停止)

不同特征基因集合的样本错分数见图 3. 从图 3 中可见, 遗传算法找到的最优染色体的基因个数为 28, 错分数为 0, 迭代次数为 3000 次.

5　胃癌预测模型规则集的抽取

本文采用自上而下的递归方法构造决策树, 包括建树和剪枝两个步骤. 并采用智能阈值搜索和剪枝策略以降低离散化数据的额外开销. 见算法 2.

算法 2

(1) Set Root node

(2) If samples are normal, return Root(返回单节点树)

(3) If samples are cancer, return Root (返回单节点树)

(4) If attribute is empty, return Root (返回规则集)

(5) Else

(6) Choose A from Attribute s(A 属性集合中分类能力最好的属性)

(7) Root classified by A

(8) For each attribute (i) in attributes

For each attributes(j) in samples(遍历每个样本中每个基因属性的值)

Set A= v_{ij}(分别以样本中的每个基因的表达值作为分类阈值)

令 v_i 为样本中满足A 的属性子集

If expression value of i is smaller than the threshold Then tree pruning (如果基因表达值小于阈值, 就剪掉这个规则分枝)

If v_i is empty (如果基因属性集合为空)

Add leaf on the parent node (在父节点下加一个叶子节点名称)

Else

Add new branch under parent node; (在父节点下加一个新的分枝)

(9) End

(10) Return root (返回预测模型的规则集)

在 28 个特征基因集合上利用决策树提取胃癌预测规则集. 首先把 40 个样本集合随机的划分成 4 组相同大小的样本子集, 每组由 5 个肿瘤样本和 5 个正常样本组成且各个子集之间没有重复基因, 以其中任意的 3 组子集作为训练集, 剩余的一组作为测试样本对该决策树的分类性能进行检验, 对每组训练集利用决策树算法对其训练, 在训练集上产生决策树, 如此反复进行 4 次, 以保证每个样本都有一次机会作为测试样本, 记录每次测试的错误率, 从而产生了 4 个不同的决策树. 如此进行多次交叉验证实验, 最终得到一个/ 决策树群0. 以每颗决策树对测试样本的分类正确率作为评价指标, 选取错误率最小的决策树提取规则集, 决策树评价结果见表 3.

表 2 中 Tr2 的分类正确率最高, 因此选取这颗决策树提取规则集. Tr2 的决策树图如图 4 所示, 图中 P 表示基因上调表达, A

表2　决策树验证结果

特征基因集合	决策树	测试正确率
COLIA2	Tr1	80%
COLIA2, ATP4B	Tr2	100%
ATP4A	Tr3	80%
SIAT1	Tr4	70%

表示基因下调表达.

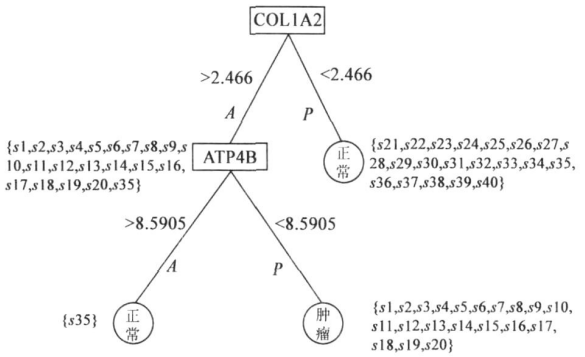

图 4 具有最大预测能力的决策树 Tr2

从决策树 Tr2 抽取的规则集描述如下:

规则集
IF COLIA2< 2.466 THEN 正常
IF COLIA2> 2.466 THEN
IF ATP4B> 8.5905 THEN 正常
IF COLIA2> 21466 THEN
I F ATP4B< 8.5905 THEN 胃癌

用上述规则集对测试样本进行类别判断, 全部样本被正确分类.

6 结论

预测模型涉及 2 个基因: COLIA2 和 ATP4B, COL1A2 基因在 20 个胃癌样本中上调表达, 在 20 个正常样本中下调表达. ATP4B 在 20 个胃癌样本中下调表达, 20 个正常样本中上调表达. 当 CIL1A2 基因在样本中的表达量超过 21466 且 ATP4B 表达量小于 815905 时, 样本为胃癌. 当 CIL1A2 基因在样本中的表达量小于 21466 时, 样本为正常.

为了进一步验证这两个基因同胃癌的关系, 我们查阅了有关文献, COL1A2 与胃癌的发生密切相关, 它是胃癌的标志基因[10] 之一. 这与本文的结论是一致. 就实验结果来看, 本文建立的胃癌分子预测模型是行之有效的.

参考文献:

[1] Kang H C, Kin I J, Park J H, et al. Identification of genes with different expression in acquired drug2resistant gastric cancer cells using high2density oligonucleotide microarrays[J]. Clinical Cancer Res, 2004, 10(1pt 1): 272- 284.

[2] Chen X, Leung S Y, Yuan S T, et al. Variation in gene expres2sion patterns in human gastric cancers[J]. Molecular biology of the cell, 2003, 14(8): 3208- 3215.

[3] Sakakura C, Hagiwara A, Nakanishi M, et al. Different gene ex2pression profiles of gastric cancer cells established from primary tumor and malignant asctites[J]. Br J Cancer, 2002, 87(10): 1153- 1161.

[4] Leung S Y, Chen X, Chu K M, et al. Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with pro2longed survival and less frequent metastasis[J]. PNAS, 2002, 99(27): 16203- 16208.

[5] Golub T R, Slonim D K, Tamayo P, et al. Molecular classifica2tion of cancer: class discovery and class prediction by gene ex2pression monitoring[J]. Science, 1999, 286(5439): 531- 537.

[6] Hippo Y, Taniguchi H, Tsutsumi S, et al. Global gene expres2sion analysis of gastric cancer by oligonucleotide microarrays. Cancer research[J]. 2002, 62(1): 233- 240.

[7] Boussioutas A, Li H, Liu J, Waring P, Lade S, Holloway A J, et al. Distinctive patterns of gene expression in premalignant gas2tric cancer[J]. Cancer research, 2003, 63(10): 2569- 2577.

[8] Pudll P, et al. Floating search methods in feature selection[J]. Pattern Recognition Letters, 1994, 15(11): 1119- 1125.

[9] Vapnik V N. The Nature of Statistical Learning Theory[M]. Springer2Verlag, 1994.

[10] Wataru Yasui, Naohide Oue, Reiko Ito, et al. Search for new biomarkers of gastric cancer through serial analysis of gene ex2pression and its clinical implications[J]. Cancer Sci, 2004, 95(5): 385- 392.

作者简介:

李 辉 男, 1971 年出生于吉林省吉林市, 1994 毕业获吉林大学学士学位, 2004 年获吉林大学硕士学位, 现为北京工业大学计算机学院博士研究生. 主要研究方向: 计算机应用技术、模式识别. E2mail: hli@emails. bjut. edu. cn

王金莲 女, 1969 年出生于陕西省千阳县, 1991 年获安徽理工大学学士学位, 2003 年获西北工业大学硕士学位, 现为北京工业大学电子信息与控制学院博士研究生. 主要研究方向: 人工智能与模式识别、生物信息.