一种肿瘤基因表达数据的知识提取方法

李颖新[1], 刘全金[2], 阮晓钢[1]
(1. 北京工业大学电子信息与控制工程学院, 北京 100022; 2. 安庆师范学院物理系, 安徽安庆 246011)

摘 要: 本文以多发性骨髓瘤的基因表达数据为例, 利用数据挖掘技术, 提出了一种针对基因表达数据进行知识发现的方法. 该方法通过计算基因的信息增益, 结合神经网络, 找出了特征基因集合, 最后利用决策树进行特征规则的提取, 给出了基于多发性骨髓瘤数据样本的产生式规则, 为生物医学研究提供了一种分析和研究基因表达数据的参考方法. 实验结果表明了该方法的有效性.

关键词: DNA 芯片; 基因表达; 数据挖掘; 神经网络; 多发性骨髓瘤
中图分类号: TP18 文献标识码: A 文章编号: 0372-2112 (2004) 09-1479-04

A Method for Extracting Knowledge from Tumor Gene Expression Data

LI Ying-xin[1], LIU Quan-jin[2], RUAN Xiao-gang[1]
(1. School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100022, China;
2. Department of Physics, Anqing Normal College, Anqing, Anhui 246011, China)

Abstract: Based on the gene expression profiles of multiple myeloma, a method was proposed for knowledge discovery using data mining and machine learning methods. We used the information gain as the criterion of each gene for classification, and got a set of informative genes using artificial neural networks. Identified by decision tree algorithm, production rules were discovered as references for biomedical researchers. The effectiveness of the method we proposed is proved by experimental results. The method can also be used as a tool for gene expression analysis in the research of biomedicine and biotechnology.

Key words: DNA chip; gene expression; data mining; neural networks; multiple myeloma

1 引言

DNA 芯片(DNA 微矩阵)技术作为一种高通量的基因表达分析平台, 可以在一次实验中同时获得成千上万个基因的表达数据. 目前 DNA 芯片技术已广泛应用于生物医学基础研究、疾病的诊断与分型以及药物筛选等研究领域[1~3]. 特别是在肿瘤学的研究中, DNA 芯片技术被越来越多的肿瘤生物学家用来分析比较肿瘤组织与对应的正常组织之间基因表达的差异, 以期发现肿瘤组织中特异表达的基因和药物治疗的靶序列. 对 DNA 芯片测定的基因表达数据进行有效地分析和建模, 从中发现肿瘤组织细胞中异常的基因表达模式已经成为生物信息学当前研究的重点. 针对 DNA 芯片的基因表达数据已有各种不同的分析方法和建模手段[4~11], 特别是数据挖掘与机器学习技术在该领域中得到了广泛的应用, 并成为了分析基因表达数据的有力工具. 但直到目前为止, 尚缺乏一种行之有效的标准方法.

对基因表达数据进行分析与建模的一个重点是在成千上万个基因中找出决定样本类别(表现型)的一组特征基因. 分析其在肿瘤组织与正常组织间的表达差异, 找出其中规律, 有助于研究肿瘤发展过程中参与的分子机制及寻找肿瘤诊断和治疗的靶分子.

本文以多发性骨髓瘤的基因表达数据为例, 利用数据挖掘与机器学习技术对该样本集进行了知识提取. 我们以基因的信息增益作为衡量基因重要性的标准, 结合人工神经网络(Artificial Neural Networks, ANN)找出了特征基因集合, 并在此基础上, 利用决策树进行知识提取. 实验结果表明了该方法的有效性.

2 问题描述

对 DNA 芯片基因表达数据进行知识提取的过程, 是一个机器学习的过程. 学习的目标是要在 DNA 芯片所测定的数千个基因中鉴别出决定样本分类的一组特征基因, 而又不丢失原始数据中包含的样本分类信息; 同时依据特征基因, 分析其中蕴含的样本分类规律, 给出形式简单、直观易懂的有关样本分类的特征规则.

我们以多发性骨髓瘤的基因表达谱为例, 进行特征基因

收稿日期: 2003-04-18; 修回日期: 2003-10-26
基金项目: 国家自然科学基金重点资助项目(No. 60234020)

与特征规则的提取. 整个知识提取系统的流程如图 1 所示.

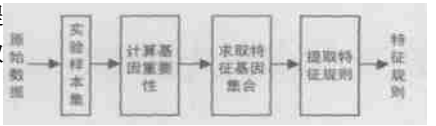

图 1 DNA 芯片基因表达数据的知识提取流程

3 实验样本集的形成

多发性骨髓瘤的 DNA 芯片基因表达数据共含 105 个样本[12], 其中 74 个为多发性骨髓瘤患者的样本数据, 31 个为正常人的样本数据. 我们排除了一个重复的患者组织样本 (p84) 后, 整个样本集共有 104 个样本, 其中患者样本数为 73, 正常样本数为 31. 每一样本均含 7129 个基因的表达数据, 表 1 是原始样本集中的一个样本的数据格式.

表 1 中/ Probe Set0 属性记录了基因探针的名称, / Avg Difference0 为基因表达水平的相对强度, / Abs Call0 属性记录了离散化后的基因表达水平值, 该属性共有 3 种可能的取值方式: A(Absent) 、P(Present) 、M(Marginal) 分别代表该基因/ 未表达0, / 表达0和/ 不能判定表达与否03 种离散化状态.

表 1 样本 P43 的数据

Experiment Name	Probe Set	Positive	Negative	Pairs	Pairs Used	Pairs In Avg	Pos Fraction	Log Avg	PM Excess	MM Excess	Pos/Neg	Avg Difference	Abs Call
P43 HuGeneFL	AFFX BioB 5_ at	9	2	20	20	18	0. 45	2. 0	0	1	4. 5	2152	P
,	,	,	,	,	,	,	,	,	,	,	,	,	,
P43 HuGeneFL	Z78285_ f_ at	0	9	20	20	19	0. 00	- 1. 33	0	0	0. 0	- 308	A

本文以离散化的基因表达水平作为分析对象, 提取每个样本文件中的/ Probe Set0 和/ Abs Call0 属性的数据, 并按表 2 的数据格式形成实验样本集.

表 2 实验样本集

SamNo	A1 (AFFX BioB 5)	A2 (AFFX BioB M)	,	A7129 (Z78285_ f)	Da (class)
s_1	A	P	,	A	myeloma
s_2	P	P	,	A	myeloma
,	,	,	,	,	,
s_{104}	P	P	,	A	normal

实验样本集共有 A1, A2, , , A7129 共 7129 个条件属性, 每个条件属性代表一个基因, 属性值为该基因的表达水平. 分类结果 Da 作为决策属性, 属性值是样本的类别/ myeloma0 或/ normal0. SamNo 属性为样本标号, 实验样本集就是样本的集合, 即 $S = \{s_1, s_2, , , s_{104}\}$.

4 基因信息增益的计算

在数据收集阶段, 对于样本分类这一任务而言, 很难确定哪些基因与此相关, 那些不相关, 因此所有基因的表达水平都被记录到样本里, 以免丢失对分类有用的信息. 然而, 在 DNA 芯片所测定的数千个基因中仅有少量基因包含了样本的分类信息, 称为/ 特征基因0. 大部分基因与样本类别并不相关, 作为噪声基因存在. 这些噪声基因不会为样本分类提供有用信息, 反而会干扰正常的知识提取.

为了评估基因对样本分类的重要性, 提供一种衡量基因分类价值的标准, 本文采用信息增益(Information Gain) 作为度量指标. 从机器学习的角度分析, 一个属性的信息增益是指由于使用了这个属性分割样本而导致的期望熵的降低. 一个属性 A 相对训练样本集合 S 的信息增益 Gain(S, A) 定义为[13]:

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

$$Entropy(S) = \sum_{i=1}^{c} - p_i \# \log_2 p_i \quad (2)$$

其中, Entropy(S) 表示样本集 S 的信息熵, values(A) 表示属性 A 的所有可能的取值, S_v 是 S 中属性 A 的值为 v 的子集. p_i 是 S 中属于类别 i 的比例. 等式(1) 中第一项 Entropy(S) 是训练样本集合 S 的熵值, 第二项是用属性 A 对样本集 S 分类后熵的期望值, 即期望熵. 这样, 就给出了由于知道属性 A 的信息而导致的熵的降低, 也就是说, 属性的信息增益反映出了由于得到了属性 A 的值后, 从而得到的关于样本分类的信息多少. 因此可以利用基因的信息熵作为其对样本分类重要性的标准.

依据各基因的信息增益, 按照从大到小的顺序将基因排序后放入有序集合 7 中:

7 = { g_1, g_2, g_3, , , g_{7129}} .

7 满足: Gain (S, g_i) E Gain(S, g_j) 且 i F j .

7 中的基因与其信息增益的关系曲线见图 2. 由图可知: 只有少数基因才具有较大的信息增益, 这也说明只有少数基因才同样本分类相关.

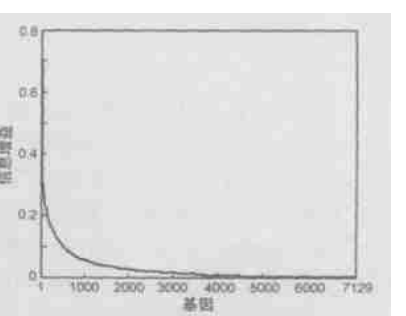

图 2 有序集合 7 中基因的信息增益曲线

5 基于人工神经网络求取特征基因集合

本文规定的特征基因集合 8, 必须同时满足以下 3 个条件:

(1) 利用该集合可对样本集 S 完全正确分类;

(2) 含有的基因数目最少;

(3) 所含基因的信息增益最大.

为此, 我们采用算法 521 和算法 522 求取特征基因集合.

算法 521 说明: 7 是按照基因的重要性由大到小排序的属性集合, 我们从中选取重要性最大的基因依次加入到有序

集合 P 中, 作为备选的特征基因集合, 根据算法 522 计算 P 作为特征基因集合时的分类误差. 通过误差与 P 的关系曲线, 得到满足上述条件的特征属性集 8 .

算法 521: GetReduSet

step1: m: = 0, P: = 5 , N: = 120

{P 是要提取的基因子集}

step2: m: = m + 1

step3: P: = PG {g_m}, 则 P 满足

P H 7 = {g_1, g_2, , , g_m}

{提取 7 中的前 m 个基因}

step4: 分类误差数 Err_m: = GetANNErr(P)

{利用神经网络误差计算算法得到 P 作为特征基因集合的分类误差数}

step5: if m F N then goto step2

step6: 依据求得的 N 个误差值: Err_1, Err_2, , , Err_n 作出误差曲线 Err2m(见图 4)

算法 522 说明: 为检查 P 作为特征基因集合时所含基因对样本的分类能力, 本文利用神经网络采用多次交叉校验的方法求取分类误差数, 并以分类误差数作为衡量 P 分类能力的标准.

算法 522: GetANNErr(P)

step1: n: = 0

step2: n: = n+ 1

step3: 将实验样本集 S 随机分成 4 组等大小的样本子集 S_1, S_2, S_3, S_4, 则这些样本子集满足如下关系:

$$\begin{cases} S_i \text{ H } S_j = 5 , \ i \text{ X } j \\ \overset{4}{\underset{i=1}{\text{G}}} S_i = S \end{cases}$$

step4: 分别以 S_1, S_2, S_3, S_4 作为 4 个不同的测试集合, 则对应的训练集为:

T_1= S- S_1, T_2= S- S_2, T_3= S- S_3, T_4= S- S_4

step5: 删除在 T_i, S_i(i = 1, 2, 3, 4) 样本中不属于集合 P 的基因. 然后将 T_i, S_i 分别作为神经网络的训练样本和测试样本. 记录所有测试样本的实际输出. 在这里, 我们采用的人工神经网络均为单隐层的 BP 网络, 网络的输入节点数 m 为 P 中基因的个数, 即 m= card(P) , 以对应 P 中各个基因的表达值, 隐层节点数为 10, 具有 1 个输出节点. 由于 P 中基因表达值有 A、M、P 三种可能的取值形式, 我们用 0、0l 5、1 与之对应.

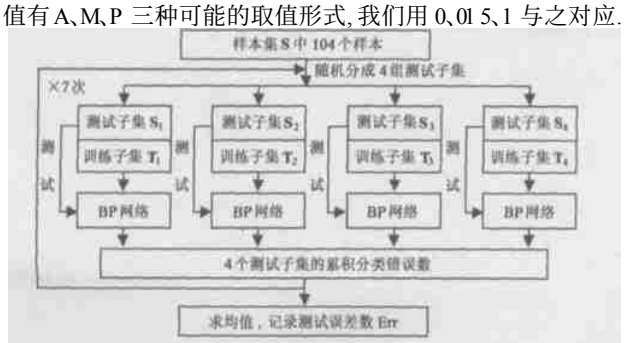

图 3　交叉校验计算分类误差算法流程

样本决策属性 Da 为 myeloma 时网络对应输出为 1, Da 取 nor2mal 时对应输出为 0.

step6: if n F 7 then goto step2

step7: 计算所有测试样本 7 次输出的平均值, 与设定阈值比较, 统计 P 作为特征基因集合时的分类误差, 并返回该误差数.

算法 522 的流程如图 3 所示.

分类误差与基因个数间的关系曲线 Err2m 见图 4. 由图 4 可知, 以 7 中前 17 个基因作为特征基因集合, 到以 7 中前 38 个基因作为特征基因集合时, 其分类误差数均为 0. 根据我们对特征基因集合的定义, 则 m= 17 时的基因子集 P 就是决定样本分类的特征基因集合 8 , 即 8 = {g_1, g_2, g_3, , , g_{17}}.

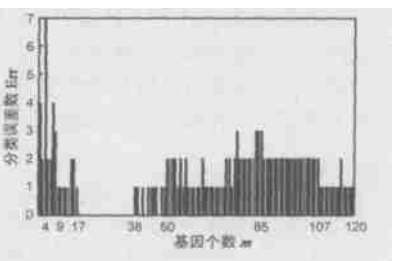

图 4　Err2m 的关系图

表 3 给出了 8 中信息增益最大的 10 个基因及其描述.

表 3　特征基因集合中信息增益最大的 10 个基因及其描述

序号	基　因	信息增益	基因描述
g_1	U24685	0l 706	GB DEF = Anti2B cell autoantibody IgM heavy chain variable V2D2J region (VH4) gene, clone
g_2	HG3872HT4142	0l 636	Immunoglobulin Gamma Heavy Chain, V(6) Djc Regions (Gb: U13200)
g_3	L00022_ s	0l 603	IG EPSILON CHAIN C REGION
g_4	HG3731HT4001	0l 510	Immunoglobulin Heavy Chain, Vdjrc Regions (Gb: L23566)
g_5	U57316	0l 508	GCN5 (hGCN5) gene
g_6	L36033	0l 481	SDF1 Stromal cell2derived factor 1
g_7	X57129	0l 461	HISTONE H1D
g_8	X57809_ s	0l 460	IGL@ Immunoglobulin lambda light chain
g_9	D59253	0l 446	NCBP interacting protein 1
g_{10}	U73167_ cds5	0l 432	SM15 gene extracted from Human cosmid LL2CA14

6　规则提取及实验结果分析

将样本集 S 中不属于特征基因集合的基因数据去掉, 则样本集仅含 17 个基因. 这样, 经过特征基因提取, 原本复杂的数据集变得简单、易于处理, 能更好地进行辅助决策. 在此样本集上, 本文采用了决策树学习算法 See5[14] 进行规则提取, 以产生式规则的形式给出了对多发性骨髓瘤基因表达数据进行规则提取的结果.

提取的规则如下:

(1) if U24685= A then myeloma(68)

(2) if U24685= M then myeloma(1)

(3) if U24685 = P then

if D59253= M then normal(0)

if D59253= P then myeloma(4)

if D59253= A then normal(31)

依据上面3 条规则, 对所有样本的错误分类数为 0. 从上述规则可以看出, 有 9311 5% 的多发性骨髓瘤的组织样本中的 U24685 未表达或低水平表达. 通过 NCBI 数据库查: U24685 为 VH4 基因. 该基因的产物为免疫球蛋白 M(IgM) 重链可变区域. Sundblad 等人研究了 IgM 与多发性骨髓瘤的关系, 证实了多发性骨髓瘤中 IgM 浓度的异常降低[15].

需要说明, 用不同方法提取的特征基因集合不同时, 对于同样的样本集提取出的规则并不会完全一致. 因此, 所提取的知识仅作为生物医学研究的参考, 为分析基因表达与疾病间的关系提供一种辅助分析手段.

7 小结

随着人类基因组计划(HGP) 的完成, 生物信息学的研究进入了后基因组时代, 利用计算方法对基因表达谱进行分析与建模是目前生物信息学研究的重点. 本文提出了一种针对离散型基因表达数据进行知识发现的方法, 重点在于鉴别出决定样本类别的特征基因, 并以此为基础, 给出数据中蕴含的样本分类知识, 从而使人们将注意力集中到可能对肿瘤形成和发展有重要关系的基因上, 做到有的放矢, 对生物医学的研究起到有益的参考作用.

应当说明的是, 验证知识发现系统所发现的信息和知识的真实性不能仅就分类正确率而定. 对其真实性的验证仍需大量的样本数据及生物学的实验.

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]. Science, 2000, 287(5459): 1777- 1782.

[3] Lander E S. Array of hope[J]. Nature Genetics, 1999, 21(suppl 1): 3- 4.

[4] Golub T R, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531- 537.

[5] Khan J, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. Nature, 2001, 7(6): 673- 679.

[6] Zhan F, et al. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance and normal bone marrow plasma cells[J]. Blood, 2002, 99(5): 1745- 1757.

[7] Yeang C H, et al. Molecular classification of multiple tumor types[J]. Bioinformatics, 2001, 17(Suppl 1): S316- S322.

[8] AliZadeh A A, et al. Distinct types of diffuse large B2cell lymphoma identified by gene expression profiling[J]. Nature, 2000, 403(6769): 503- 511.

[9] DeRisi J, et al. Use of a cDNA mircoarray to analyse gene expression patterns in human cancer[J]. Nature Genet, 1996, 14(4): 457- 460.

[10] Hamadeh H K, et al. Prediction of compound signature using high density gene expression profiling[J]. Toxicological Science, 2002, 67(2): 232- 240.

[11] Sherlock G. Analysis of large2scale gene expression data[J]. Current Opinion in Immunology, 2000, 12(2): 201- 205.

[12] Lambert Lab of myeloma genetics. Gene expression profile of multiple myeloma[DB/OL]. http://lambertlab.uams.edu/publicdata.htm, 2001.

[13] Mitchell T M. 机器学习[M]. 北京: 机械工业出版社, 2003.

[14] Quinlan J R. See5.0: RuleQuest Research Data Mining Tools[CP]. http://www.relequest.com, 2002.

[15] Sundblad A, et al. V2region2specific alterations of serum IgM production in multiple myeloma of IgG class[J]. The Hematology Journal, 2000, 1(2): 102- 110.

作者简介:

李颖新 男, 1972 年 9 月生于河北迁安, 1995 年获河北科技大学自动化系工学学士学位, 现为北京工业大学模式识别与智能系统专业博士研究生, 主要研究方向: 模式识别、人工智能、数据挖掘与机器学习、生物信息学.

刘全金 男, 1971 年 12 月出生于安徽寿县, 安徽省安庆师范学院物理系教师, 中职, 现为北京工业大学电子信息与控制工程学院访问学者, 主要研究方向: 信息处理、生物信息学.