全国第七届研究生数学建模竞赛

题 目 A　　确定肿瘤的重要基因信息

——提取基因图谱信息方法的研究

摘　　要：

对于问题一，我们首先对数据进行前期的预处理，然后分别建立评分模型，对各个基因进行打分，从而按照分数又高到低排序，然后用主成分析法，确定包含样本全部信息的基因集的基因个数，最后用因子分析的方法提取出基因中潜在的少量可以完全表达样本信息的因子，共7个，我们称之为因素。

问题二实际上就是要求我们找出能够表达全部样本的最少的基因标签的个数。我们利用REF-relief算法和支持向量机（SVM）进行选择，最后使用“留一法”对其进行误差分析，最终得到最少的基因标签为5个，分别是X63629、H06524、H08393、R39209、M26383，采用“留一法”得到最后错判的数量为6个，正确率为90.3%。

问题三实际上就是让我们在问题二的基础上，建立含有噪声的模型，对含噪声模型进行分析，看是否能发现噪声对分类是有帮助的。我们引入控制因子，提出对噪声的分析是有利于分类，如果控制因子选取在合理的范围并且方向也选取合适，那么可以大大降低样本错判的数量。

第四问就是让我们能够利用信息融合以及数据挖掘的方法，建立起一个分类算法的决策树，由于信息来自多个方面，因此对海量信息有效提取和挖掘是十分又必须的，终于我们得到了一个基于二叉树的基因诊断模型，该模型能有效地对样本进行分类，结果仅有1个样本错判。

目录

1 问题重述 3
2 模型假设 4
3 符号说明 4
4 问题分析 5
4.1 问题 1 分析 5
4.2 问题 2 分析 5
4.3 问题 3 分析 5
4.4 问题 4 分析 5
5 模型建立与求解 6
5.1 数据预处理 6
5.2 问题 1 模型与求解 7
5.3 问题 2 模型与求解 14
5.4 问题 3 模型与求解 16
5.5 问题 4 模型与求解 20
6 附录 27
7 参考文献 29

1 问题重述

癌症起源于正常组织在物理或化学致癌物的诱导下，基因组发生的突变，即基因在结构上发生碱基对的组成或排列顺序的改变，因而改变了基因原来的正常分布（即所包含基因的种类和各类基因以该基因转录的mRNA的多少来衡量的表达水平）。所以探讨基因分布的改变与癌症发生之间的关系具有深远的意义。

DNA微阵列（DNA microarray），也叫基因芯片，是最近数年发展起来的一种能快速、高效检测DNA片段序列、基因表达水平的新技术。它将数目从几百个到上百万个不等的称之为探针的核苷酸序列固定在小的（约1 cm^2）玻璃或硅片等固体基片或膜上，该固定有探针的基片就称之为DNA微阵列。根据核苷酸分子在形成双链时遵循碱基互补原则，就可以检测出样本中与探针阵列中互补的核苷酸片段，从而得到样本中关于基因表达的信息，这就是基因表达谱，因此基因表达谱可以用一个矩阵或一个向量来表示，矩阵或向量元素的数值大小即该基因的表达水平（见附件）。

随着大规模基因表达谱(Gene expression profile,或称为基因表达分布图)技术的发展，人类各种组织的正常的基因表达已经获得，各类病人的基因表达分布图都有了参考的基准，因此基因表达数据的分析与建模已经成为生物信息学研究领域中的重要课题。如果可以在分子水平上利用基因表达分布图准确地进行肿瘤亚型的识别，对诊断和治疗肿瘤具有重要意义。因为每一种肿瘤都有其基因的特征表达谱（见附图）。从DNA芯片所测量的成千上万个基因中，找出决定样本类别的一组基因“标签”，即“信息基因”（informative genes）是正确识别肿瘤类型、给出可靠诊断和简化实验分析的关键所在，同时也为抗癌药物的研制提供了捷径。

通常由于基因数目很大，在判断肿瘤基因标签的过程中，需要剔除掉大量“无关基因”，从而大大缩小需要搜索的致癌基因范围。事实上，在基因表达谱中，一些基因的表达水平在所有样本中都非常接近。例如，不少基因在急性白血病亚型（ALL,AML）两个类别中的分布无论其均值还是方差均无明显差别,可以认为这些基因与样本类别无关,没有对样本类型的判别提供有用信息,反而增加信息基因搜索的计算复杂度。因此,必须对这些“无关基因”进行剔除。1999年《Science》发表了Golub 等针对上述急性白血病亚型识别与信息基因选取问题的研究结果[1]。Golub 等以“信噪比”(Signal to noise ratio)指标作为衡量基因对样本分类贡献大小的量度,采用加权投票的方法进行亚型的识别,仅根据72个样本就从7129个基因中选出了50个可能与亚型分类相关的信息基因。Golub的工作大大缩小了决定急性白血病亚型差异的基因范围，给出了亚型识别的基因依据，富有创造性。Guyon 等则利用支持向量机的方法再从中选出了8个可能的信息基因[2]。

但信噪比肯定不是衡量基因对样本分类贡献大小的唯一标准，肿瘤是致癌基因、抑癌基因、促癌基因和蛋白质通过多种方式作用的结果，在确定某种肿瘤的基因标签时，应该设法充分利用其他有价值的信息。有专家认为[3]在基因分类研究中忽略基因低水平表达、差异不大的表达的倾向应该被纠正，与临床问题相关的主要生理学信息（见问题4）应该融合到基因分类研究中。

面对提取基因图谱信息这样前沿性课题，命题人根据自己科学研究的经历和思考，猜测以下几点是解决前沿性课题的有价值的工作。这种猜测是科学研究中

的重要环节，当然猜测不会总是可行的，更不一定总是正确的。但不探索就不能前进，如果能够通过数学建模，得到的部分结果可以佐证你们的猜测或为新探索提供若干依据，就很有价值。我们的目的只是给研究生以启发，鼓励研究生培养这样的创造性发现的能力。所以研究生完全可以独立设计自己的技术路线，只要能够有效提取附件的基因图谱信息就行。

（1） 由于基因表示之间存在着很强的相关性，所以对于某种特定的肿瘤，似乎会有大量的基因都与该肿瘤类型识别相关，但一般认为与一种肿瘤直接相关的突变基因数目很少。对于给定的数据（见附件），如何从上述观点出发，选择最好的分类因素？

（2） 相对于基因数目，样本往往很小，如果直接用于分类会造成小样本的学习问题，如何减少用于分类识别的基因特征是分类问题的核心，事实上只有当这种特征较少时，分类的效果才更好些。对于给定的结肠癌数据如何从分类的角度确定相应的基因“标签”？

（3） 基因表达谱中不可避免地含有噪声（见 1999 年 Golub 在《Science》发表的文章），有的噪声强度甚至较大，对含有噪声的基因表达谱提取信息时会产生偏差。通过建立噪声模型，分析给定数据中的噪声能否对确定基因标签产生有利的影响？

（4） 在肿瘤研究领域通常会已知若干个信息基因与某种癌症的关系密切，建立融入了这些有助于诊断肿瘤信息的确定基因“标签”的数学模型。比如临床有下面的生理学信息：大约 90%结肠癌在早期有 5 号染色体长臂 APC 基因的失活，而只有 40%~50%的 ras 相关基因突变。

2 模型假设

（1）假设数据是真实可信，没有任何人为主观因素的干扰，即不存在人为的数据修正，能够反映基因序列的真实含义。

（2）假设基因不存在缺失或者未转录成功的情况。

（3）假设不存在噪声和任何干扰的情况下，总存在最优的分类方案。

（4）第一问和第二问均在假设数据无误差的情况下建立的，第三问是在假设数据有噪声的情况下建立的。

3 符号说明

设 $G=\{g_1,g_2,...,g_n\}$ 表示一个DNA序列中所有基因组成的一个基因集合，其中 $g_i(1\le i\le n)$ 表示一个基因，$|G|=n$ 表示全部基因的个数。设 $G=\{s_1,s_2,...,s_n\}$ 表示由实验样本构成的样本集合，其中 $|S|=m$ 表示样本数量，每一个样本 $s_i(1\le i\le m)$ 表示在某种条件下一个DNA序列中所有基因的表达式，即 $s_i(1\le i\le m)$ 是一个n维空间向量，且 $s_i\in R^n$。由所有的样本及其所属类别所组成的基因表达矩阵M可表示为

$$M=,n\left[\begin{array}{cccc|c} \overbrace{x_{1,1} & x_{1,2} & \cdots & x_{1,n}}^{n\text{个基因}} & \overbrace{l_1}^{\text{类别}} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} & l_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} & l_m \end{array}\right]$$

其中 $x_{i,j}$ 表示基因 g_j 在样本 s_i 中的基因表达值，通常情况下 $m \le n$。矩阵M的一行表示同一条件下所有基因的基因表达谱，构成n维空间的一个点，而矩阵M的一列表示一个基因在不同条件下的表达情况。Class表示相应样本的所属类别，实际就是指一个样本是属于正常样本类还是属于肿瘤组织样本类。

4 问题分析

本题涉及到生物学的基因这一前沿课题，主要要求参赛者具有一定的分类和信息融合的知识，该题目从不同角度要求参赛者利用所给的数据进行有效信息的提取，由于数据量很大，并且冗余信息也很多，这就要求参赛者能够利用各种有效的方法提取出必要的信息，从而删除相关度较高的一部分信息，因为相关度高的数据之间不仅使得基因量远远大于样本，并且里面的噪声也会对分类产生一定的影响，但究竟如何提取有效的信息，是第一问需要解决的问题；另外，第二问从另一个角度对该问题进行描述，实际要求我们能够从给定的几千个基因中找出特征基因，即就是能够找到一组数量少，并且能够涵盖原来所有信息的基因组；问题三是从噪声的角度，考虑分析噪声能够对我们分类产生有利的影响，因为噪声的存在肯定会带来影响，但究竟是好的还是坏的，就看我们如何利用；问题四实际是一个信息融合和数据挖掘的问题，要求我们能够建立很好的决策树，方便对未知的样本进行分类。

4.1 问题 1 分析

对于问题一，我们首先对数据进行前期的预处理，然后分别建立记分模型，对各个基因进行打分，从而按照分数又高到低排序，然后用主成分析法，确定包含样本全部信息的基因集的基因个数，最后用因子分析的方法提取出基因中潜在的少量可以完全表达样本信息的因子，我们称之为因素。

4.2 问题 2 分析

该问题实际上就是要求我们找出能够表达全部样本的最少的基因标签的个数。我们利用REF-relief算法和支持向量机（SVM）进行选择，最后使用“留一法”对其进行误差分析。

4.3 问题 3 分析

第三问实际上就是让我们在以前的基础上，建立含有噪声的模型，对含噪声模型进行分析，看是否能发现噪声对分类是有帮助的。

4.4 问题 4 分析

该问题就是让我们能够利用信息融合以及数据挖掘的方法，建立起一个分类算法的决策树，由于信息来自多个方面，因此对海量信息有效提取和挖掘是十分又必须的。

5 模型建立与求解

5.1 数据预处理

对于肿瘤的分类与识别，都是以大规模基因表达数据作为分析的基础。由于样本获取和制备上的困难，因此目前的基因表达谱的样本数很少。这就造成了基因表达谱样本少，维数高，噪声大的特点。在对基因表达数据进行聚类、分类等数据分析之前，往往需要进行预处理，才可能得到反映生物本质的分类和聚类的结果。当然基因表达谱的预处理过程包括很多部分，例如对丢失数据进行填补、清除不完整的数据或合并重复数据等数据清洗，根据分析的目的进行数据过滤，以针对不同的分析方法选择合适的数据转换等。由假设可知道，数据存在丢失或者不完全的信息，因此，必须对数据进行合并重复数据的数据，主要是project_data1.xls文件中来自相同ESTs的数据进行合并。基因表达谱在数据过滤前，往往还需要进行数据转换。数据转换就是将数据变换为适合数据挖掘的形式，可以根据需要构造出新的数据属性以帮助理解、分析数据的特点，或者将数据规范化，使之落在一个特定的数据区间中。数据转换主要有以下四种可选操作类型：对数转换，归一化，均值中心化和中间值中间化。

（1） 对数变换

为了反映某个基因表达水平在实验样本和参考样本中的对数关系，要对表达值进行对数变换。常用的对数底为2、e、10等。考虑时间序列上的基因表达谱，实验结果是相对于0时刻的表达水平。举例来说，设第一个点的值为1.0，第二个点的值为第一个点值的E调二倍，第三个点是下调二倍，那么这三个点的值应该依次为：1，2，0.5；其实第二点，第三点较之第一点都是两倍的幅度变化关系，只不过变化方向不同，从这三个原始值中却不太容易看出这一点来。如果对这两个值取以2为底的对数，得到0，1，-1。从中我们很容易看出第二点和第三点相对于第一点成二倍对称关系。因此，因此数据转换可以使小于1的值变大，大于1的值变小，从而使他们关于0对称化，这种变换反映了一定的生物学意思，能更直观地了解基因地上调和下调幅度。而附件中所给的excel文件经过了$\log_2$变换。

（2） 均值/中值的中心变化

由基因芯片试验原理和过程可知，在各个组织样本实验水平下，得到的是一系列经过$\log_2$变换的数据。然而，样本本身对于我们的研究没有意义，为了使数据不依赖于参考值，需要对数据进行调整，使其仅仅反映相对于观察值总体(如均值、中值)的变化，通过均值或中值的中心变化可以做到这一点，我们称之为标准化变换。另外，对列进行中心化还可以在一定程度上消除某些类型的系统误差。这些系统偏差主要来源于杂交实验中的RNA容量差异、标记效率和图像扫描参数的不同。另外，经研究发现，最好在对数空间进行标准化操作效果更好。

假设 $\widetilde{x_{ik}} = \dfrac{x_{ik} - \mu_k}{\sigma_k}$， i=1,2,…m; k=1,2,…p。其中 $\widetilde{x_{ik}}$ 表示修正后的数据，x_{ik} 表示第i个样本的未做变换值，μ 表示所有样本的均值，σ 表示所有样本的方差。并且有 $\mu_k = \dfrac{1}{m}\sum_{i=1}^{m} x_{ik}$，$\sigma_k^2 = \dfrac{1}{m-1}\sum_{i=1}^{m}(x_{ik} - \mu_k)^2$。目的是消除量纲对评价结果的影响，得到标准化后的矩阵。

（3） 数据合并

将project_data1.xls中来自相同ESTs的数据进行合并。合并的方式有很多，我们这里面采用最简单的求取平均值。

结论：最终我们得到了，原文件中一共包含有1911个不同的基因片段。

5.2 问题 1 模型与求解

5.2.1 理论分析

由于基因表示之间存在着很强的相关性，所以对于某种特定的肿瘤，似乎会有大量的基因都与该肿瘤类型识别相关，但一般认为与一种肿瘤直接相关的突变基因数目很少。对于给定的数据（见附件），如何从上述观点出发，选择最好的分类因素。

将数据进行上述的（1）->（2）->（3）变换，可以得到最后修正后的数据，之后我们都将使用修正后的数据进行分析。

5.2.2 模型建立

第一步　记分准则建立

由于基因表达谱存在维数高、噪音大以及冗余信息多等特点，所以在分类前需要采用各种方法对基因表达谱数据进行降维、去噪和剔除冗余基因等处理，以最大限度地提高肿瘤样本的分类性能．基因表达谱中多数基因的表达与肿瘤无关，这为信息基因选择带来很大困难，采用单一方法很难完成这一任务，因此，通常首先采用一种称之为基因排序(Gene Ranking)的方法对原始基因集合进行粗选，其基本思想便是按照某种记分准则对每一个基因进行记分，基因分值大小反映了基因的重要程度，然后按基因分值大小降序排列基因并选择排在前面的一定数量的基因作为选择结果。

Golub等人采用特征记分准则(Feature ScoreCriterion，FSC)对每一个基因计算其分值，然后按分值降序排列基因，基因的分值通过式(1)计算：

$$FSC(g_i)=\left(\mu_i^+-\mu_i^-\right)\Big/\left(\delta_i^+-\delta_i^-\right) \tag{1}$$

其中，μ_i^+ 表示基因 g_i 的正类样本的均值，μ_i^- 表示 g_i 的负类样本的均值；而 δ_i^+ 表示 g_i 的正类样本的标准差，δ_i^- 表示 g_i 的负类样本的标准差，基因分值的大小表明该基因能够区分正类样本和负类样本的分类能力大小。李颖新等认为在衡量基因含有样本分类信息多少的度量问题上，还应该考虑由于方差不同所带来的对样本分类的贡献，从而可以更客观地评价基因含有的分类信息量，为此他们对FSC进行了修订，称之为修订的特征记分准则(Revised Fea—ture Score Criterion，RFSC)，如式(5)所示：

$$FSC(g_i)=0.5\times\left|\left(\mu_i^+-\mu_i^-\right)\Big/\left(\delta_i^+-\delta_i^-\right)\right|+0.5\times\ln\left(\left(\left(\delta_i^+\right)^2+\left(\delta_i^-\right)^2\right)\Big/2\delta_i^+\delta_i^-\right) \tag{2}$$

更进一步，如果假设两类样本的分布都服从高斯分布，则可根据基因 g_i 采用Bhattacharyya距离作为两类样本的可分性判据，称之为Bhatta—charyya特征记分准则(Bhattacharyya Feature Score Criterion，BFSC)，即以式(6)来度量基因 g_i 的分类能力。

$$FSC(g_i)=1/4\times\left(\mu_i^+-\mu_i^-\right)^2\Big/\left(\left(\delta_i^+\right)^2+\left(\delta_i^-\right)^2\right)+0.5\times\ln\left(\left(\left(\delta_i^+\right)^2+\left(\delta_i^-\right)^2\right)\Big/2\delta_i^+\delta_i^-\right) \tag{3}$$

本文将同时对三种分类准则进行研究，带着探索的目的，尝试得到适合本问题的最好的分类准则。由于基因的数据量非常多，因此我们只列取排名前20的基因。

表1　判别准则表

排名顺序	FSC准则		RFSC准则		BFSC准则	
	基因名称	FSC值	基因名称	RFSC值	基因名称	BFSC值
1	R87126	0.8347	R87126	0.4184	R87126	0.3486
2	R36977	0.7389	R36977	0.3754	R36977	0.2752
3	H08393	0.7241	M63391	0.3687	H08393	0.2622
4	M26383	0.7228	M26383	0.3628	M26383	0.2618
5	M63391	0.7106	H08393	0.3621	M63391	0.2583
6	X12671	0.6966	M22382	0.3571	X12671	0.2440
7	M22382	0.6881	X12671	0.3514	M22382	0.2428
8	R44887	0.6878	R44887	0.3460	R44887	0.2375
9	Z50753	0.6807	J05032	0.3427	Z50753	0.2319
10	J05032	0.6676	Z50753	0.3407	J05032	0.2273
11	X63629	0.6660	X63629	0.3367	X63629	0.2236
12	H43887	0.6559	M76378	0.3362	H43887	0.2167
13	R14852	0.6336	H43887	0.3311	M76378	0.2105
14	M76378	0.6298	U30825	0.3297	R14852	0.2023
15	T88712	0.6203	R14852	0.3196	U30825	0.1981
· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·
1905	V00523	0.0000	V00523	0.0013	T86281	0.0000
1096	J03600	0.0000	J03600	0.0011	R54837	0.0000
1097	X56253	0.0000	X56253	0.0000	M99626	0.0000
1908	J02906	0.0000	J02906	0.0000	R85326	0.0000
1909	H20543	0.0000	H20543	0.0000	T56674	0.0000
1910	H08144	0.0000	H08144	0.0000	M59819	0.0000
1911	V00520	0.0000	V00520	0.0000	R54854	0.0000

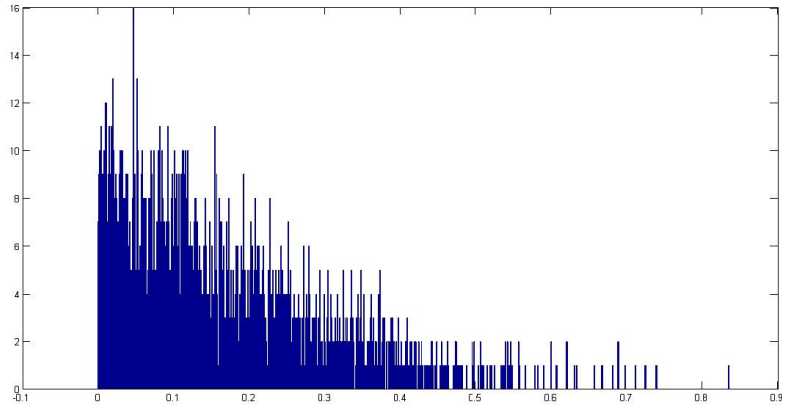


图1　FSC

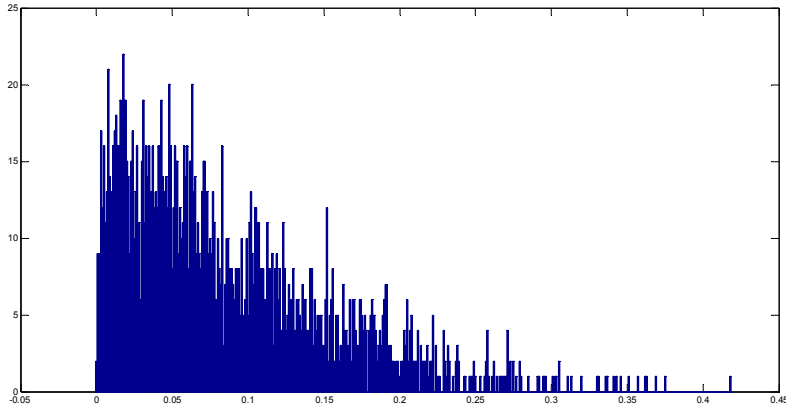


图2 RFSC

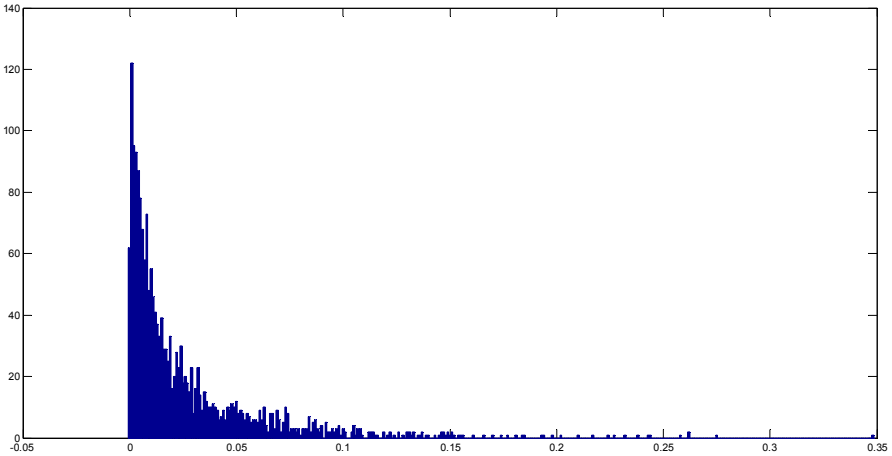


图3 BFSC

第二步 主成分析法

PCA是多元统计分析中应用广泛的一种数据降维方法，是一种重要的掌握事物主要矛盾的统计分析方法，它是研究如何通过原始变量的少数几个线性组合来解释多变量的方差，它可以从多元事物中解析出主要影响因素，揭示事物的本质，简化复杂的问题。计算主成份的目的是将高维数据映射到较低维空间。PCA的目标是寻找：r个新变量，使它们反映事物的主要特征，压缩原有数据矩阵的规模。每个新变量是原有变量的线性组合，体现原有变量的综合效果，具有一定的实际含义。这r个新变量称为主成份分量，它们可以在很大程度上反映原来n个变量的影响，并且这些新变量是互不相关的，也是正交的。通过主成份分析，压缩数据空间，将多元数据的特征在低维空间里直观地表示出来。

抽取主成分析分量的基本过程可以描述为：对降维后的矩阵 M 、进行主成份分析并从中抽取主成份分量。为使样本集 M 在降维过程中所引起的平方误差最小，必须进行两方面的工作：一是进行坐标变
换，即用雅可比方法求解正交变换矩阵;二是选取w个主成份分量，w<p。PCA的计算过程主要分三步进行。

1．用标准化后的样本矩阵进行分析，并计算矩阵 $\widetilde{M}$ 的相关系数矩阵R。

2．对于相关系数矩阵R，采用雅克比方法求特征方程 $|R-\lambda I|=0$ 的 p 个非负的特征值 $\lambda_1>\lambda_2>\ldots>\lambda_p\geq 0$ ， λ_i 的特征向量为 $v_i=(v_{i1},v_{i2},\ldots v_{ip})$ ， i=1,2,…p，并且满足 $v_iv_j=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}$ 。

3．选择w个主成份分量，使得前面w个主成份的方差和占全部总方差的比例 $\eta=\sum_{i=1}^{w}\lambda_i\Big/\sum_{j=1}^{p}\lambda_j$ 接近1,并使所选的这w个主成份尽可能多地保留原来p个基因的信息，从而达到降维的目的，得到主成份矩阵 $\widetilde{M_w}$，$\widetilde{M_w}$ 中的样本类别信息仍保持。

通过实验仿真，我们采用了两类数据集，即正常数据和肿瘤数据。我们不妨定义正常数据为正类样本，肿瘤数据为负类样本。附录数据给出了62个样本，其中正类样本为22个，负类样本为40个。这样我们可以用一个矩阵 $M=(x_{ij})_{62\times1911}$ 来表示。利用第一步的记分准则RFSC的方法，对原来的基因按照得分进行排序，我们分别取前面的200、150、75个基因继续进行PCA的训练。

从而得到如下表：

表2　主成份分析与贡献度表

基因排序	基因标签	贡献度 $\eta=\sum_{i=1}^{w}\lambda_i\Big/\sum_{j=1}^{p}\lambda_j$
1	R87126	0.4715
2	R36977	0.5616
3	M63391	0.6185
4	M26383	0.6735
5	H08393	0.7184
6	M22382	0.7500
7	X12671	0.7771
8	R44887	0.7957
9	J05032	0.8098
10	Z50753	0.8219
.	.	.
.	.	.
.	.	.
55	T61609	0.9931
56	R64115	0.9944
57	T58861	0.9956
58	T57468	0.9968
59	H20426	0.9980
60	U22055	0.9990
61	D38551	1.0000
62	R10066	1.0000
63	T52185	1.0000
.	.	.
.	.	.
.	.	.

从上表中，我们可以得出，前面61个基因的贡献度已经为 $\eta=\sum_{i=1}^{w}\lambda_i\Big/\sum_{j=1}^{p}\lambda_j$ =1，这说明了这前61个基因能够完全表达所有的基因的信息。附录给出了这61个基因

的标号和排列顺序，其中排在前面的基因用RFSC得到的得分越大。
结论：因此我们找到了61组基因来刻画整个基因样本集。

第三步　因子分析法

我们可以发现一共有62个样本，而对这61个基因进行分析，由于分类要将这61个样本分成两部分，其中一部分为训练集，另一部分为测试集；那么显然可以发现基因量还是大于训练样本的数量，从另一个角度讲，虽然采用主成份分析方法从基因表达谱中提取主成分实现了数据降维，而且降维后的分类实验结果也令人满意，但是所抽取的主成份缺少生物学含义，因此有必要进一步采用因子分析方法来做同样的实验，因为针对具体问题因子分析方法中的因子具有可解释性，更重的是，能够从整个基因样本中发现一些不能直接从样本基因从得出的潜在的因素，所以，如何进一步减少基因数量，提取有效成分，是十分重要的。

因子分析方法与主成份分析方法是两种分析实值随机变量相关结构的不同分析方法。因子分析能够揭示一组观测变量的潜在因素，因此，它能够从高维样本中发现影响目标对象观测值的潜在因子，从而降低观测样本的维数。因子分析的主要目标就是判定影响观测值的公共因子数量以及因子与观测值之间的联系程度。因子分析主要有两种形式:探索性因子分析(ExPloratory factor analysis，EFA)和验证性因子分析(Confirmatoryfactoranalysis，CFA)。EFA是一种用来找出多元观测变量的本质结构并进行数据降维的技术，因而，EFA能够将具有错综复杂关系的变量综合为少数几个核心因子;CFA则用于测试一个结构是否正影响观测变量。

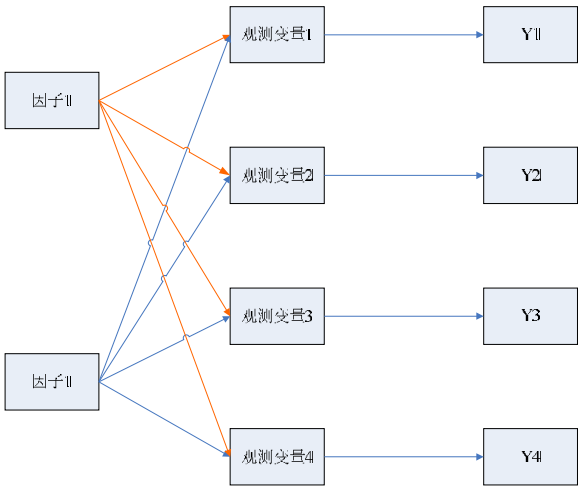


图4 因子分析示意图

设 $x=(x_1,x_2,...x_n)'$ 表示观测向量，它的均值向量记为 $\mu=(\mu_1,\mu_2,...\mu_n)'$，所以因子模型可以表示为 $x=\mu+Af+\varepsilon$，其中 $f=(f_1,f_2,...f_n)'$ $(r\le m)$ 表示公共因子向量，$\varepsilon=(\varepsilon_1,\varepsilon_2,...\varepsilon_n)'$ 表示因子误差向量，$A=(a_{ij})_{m\times r}$ 表示因子载荷矩阵。首先根据Kaiser准则决定因子最佳数量，因子数量应该与相关矩阵的大于1的特征值的数量相等。然后根据主成分析方法抽取因子的初始集，再借助Varimax正交旋转方法旋转所得到的因子，所以每一个测量值都与各因子线性相关。其相关程度可以从因子载荷体现出来，这个载荷矩阵可解释为标准化了的回归系数。

(1) 选择分析的变量

选择分析的变量用定性分析和定量分析的方法选择变量，因子分析的前提条件是观测变量间有较强的相关性，因为如果变量之间无相关性或相关性较小的话，他们不会有共享因子,所以原始变量间应该有较强的相关性。

(2) 计算所选原始变量的相关系数矩阵

相关系数矩阵描述了原始变量之间的相关关系。可以帮助判断原始变量之间是否存在相关关系，这对因子分析是非常重要的，因为如果所选变量之间无关系，做因子分析是不恰当的。并且相关系数矩阵是估计因子结构的基础。

(3) 提取公共因子

这一步要确定因子求解的方法和因子的个数。需要根据研究者的设计方案或有关的经验或知识事先确定。因子个数的确定可以根据因子方差的大小。只取方差大于1(或特征值大于1)的那些因子，因为方差小于1的因子其贡献可能很小；按照因子的累计方差贡献率来确定，一般认为要达到85%才能符合要求；

(4) 因子旋转

通过坐标变换使每个原始变量在尽可能少的因子之间有密切的关系，这样因子解的实际意义更容易解释,并为每个潜在因子赋予有实际意义的名字。

(5)计算因子得分

求出各样本的因子得分，有了因子得分值，则可以在许多分析中使用这些因子，例如以因子的得分做聚类分析的变量，做回归分析中的回归因子。

5.2.3 仿真实例

第一种方法，我们采用R型因子分析法，利用上面主成分析法得出的61个基因进行进一步的分类。

第一步可得到相关矩阵的特征值为

表3 相关矩阵与特征值表

序号	特征值
1	32.61959581179523
2	8.349776779804627
3	4.600997123480855
4	2.257141681610037
5	1.968187079167144
6	1.210923803225692
7	**1.152337491050264**
8	0.885589196506306
9	0.842097456852102
. . .	. . .
59	0.000216261053578
60	0.000075079104095
61	0.000005155023659

根据上述结论3，我们可以知道，这61个基因组包含了7个重要的因素。我们不妨把这7个重要因素定义为 F_1， F_2， F_3， F_4， F_5， F_6， F_7。并且这7个主因子的

贡献度为：

表4 因子分析与贡献率表

主因子	贡献率
F_1	20.50502683017677
F_2	8.693274617330005
F_3	11.421048588701233
F_4	6.471791464888556
F_5	1.656662117449377
F_6	1.874006892295331
F_7	1.537149259292555

另外我们可以得到因子得分函数和R-型的得分，得分矩阵为

表5 因子得分表

因子 / 得分	F_1	F_2	F_3	F_4	F_5	F_6	F_7
1	-0.9170	-1.4840	-0.8585	0.5721	-0.5925	0.7056	0.8695
2	-1.4362	0.1180	-0.6528	-0.3468	-1.7157	0.4904	1.6277
3	-0.3196	-0.2991	2.1435	-0.7899	-1.2569	0.4894	-0.4148
4	-0.3708	-0.4860	1.9369	-0.1545	-0.6588	-0.2707	0.5580
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
61	1.4906	0.1196	-0.3815	-0.9969	2.0762	0.5853	1.2381
62	-0.3138	0.9585	-0.1904	-0.0358	1.5125	-1.4841	0.5617

很多文献都会提出用F因子贡献度比较大两个因素作为隐含因子，但是我们可以做出2维图，以 F_1 和 F_3 的得分作为分类，我们可以看出

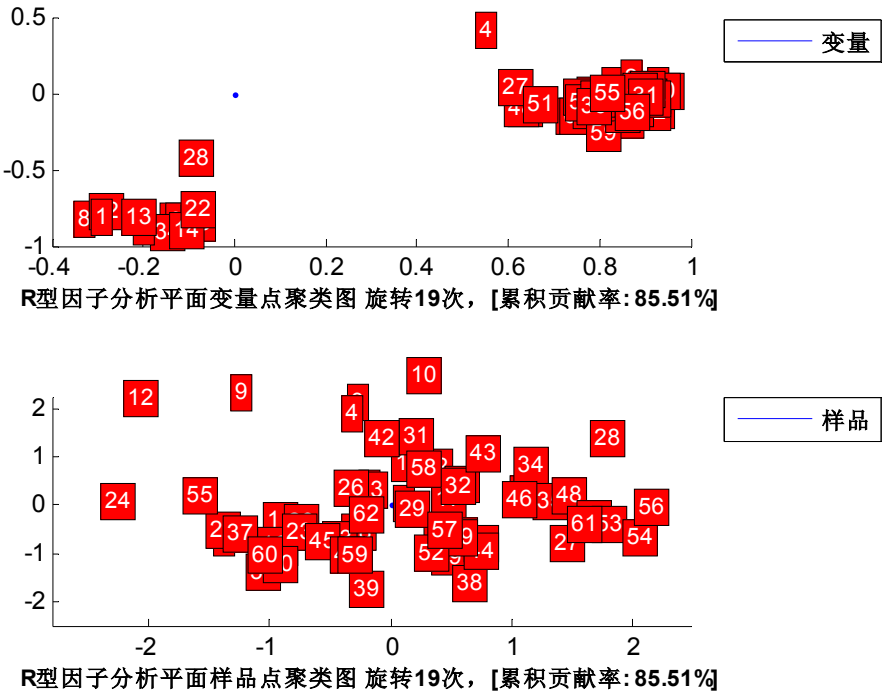


图5 两主因子分布图

上图说明，如果仅以2个主导因子显然是不能包含全部信息，因此这7个因子共同影响着基因。

5.2.4 结论

最终我们找出影基因选择最好的分类因素有 7 个，并且这七个因素影响着主成分析得到的 61 个主导基因，从而影响整个分类的结果，最后我们可以以这 7 个因素的得分矩阵进行聚类分析。

另外，我们找到了 7 个因素，但是如何根据这七个因素分类，并且分类效果好不好是有待检验，这就引出了第二问的分析解答。

5.3 问题 2 模型与求解

5.3.1 理论分析

相对于基因数目，样本往往很小，如果直接用于分类会造成小样本的学习问题，如何减少用于分类识别的基因特征是分类问题的核心，事实上只有当这种特征较少时，分类的效果才更好些。对于给定的结肠癌数据如何从分类的角度确定相应的基因“标签”。

5.3.2 模型建立

基于上述分析，本文在借鉴递归特征排除(recursive feature elimination, RFE)策略思想的基础上采用如下过程进行样本分类特征的选取：首先，对当前属性集合F中的所有属性利用Relief算法进行属性重要性的排序，去掉具有最小分类权重的那个属性；然后，重新采用Relief算法计算剩余属性的分类权重，再排除这些属性中具有最小权重的属性… 如此循环下去，就使得噪声属性被逐步剔除，Relief算法对属性分类能力的评价受噪声属性的影响将不断减小，得到的属性分类权重也就越接近真实。该算法在本文中称为RFE_Relief算法：

A. RFE-Relief algorithm:

(1) F:={$g_1, g_1,...g_p$}, n:=0

(2) while $F \neq \varnothing$ do

(a) $F_n := F$, n:=n+1 {记录当前集合}

(b)W:=Relief(F); {利用Relief算法计算当前属性集合F中属性的分类权重}

(c) c:=arg min W {找到具有最小权重属性的位置}

(d) F:=F-F(c) {从属性集合中去除该属性}

B. Relief algorithm:

Relief 算法(S_{trn}). //F 为待分析的属性集合，S_{trn} 为训练样本集

1．set weights vector W to zeros

//向量 W 中第 i 个元素对应于 F 的第 i 个属性的分类权重

2．For i=1 to card(S_{trn})

//card(S_{trn})为样本集 S_{trn} 中的样本数

2.1 choose i-th instance s_i in S_{trn}

2.2 Find its nearest K Hits nearest K Misses

//$K \geq 1$, $K > 1$时为 Relief-A算法

2.3 For j=1 to card(S_{trn})

$$W_j = W_j - \frac{\sum_{m=1}^{K}\left(s_{ij} - nearestHit_{mj}\right)^2}{K} + \frac{\sum_{m=1}^{K}\left(s_{ij} - nearestMiss_{mj}\right)^2}{K}$$

3． Return W //返回权值向量

C. 支持向量机 SVM

支持向量机(support vector machine, SVM)是由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一种机器学习算法，可在有限样本的条件下获得良好的推广能力.若给定样本集为:

$S_T = \left\{(x_i, y_i) \mid x_i \in R^d, y_i \in \{-1, +1\}, i = 1, 2, ..., N\right\}$, 在线性可分条件下SVM将该样本集的样本分类问题转化为如下二次规划问题的求解。

$$\text{minimize:}\quad \Phi(w) = \frac{1}{2}\|w\|^2$$

$$\text{subject to:}\quad y_i(w^T x_i + w_0) \geq 1, i = 1, 2, ..., N$$

其二次对偶问题可以表示为：

$$\begin{cases} \max \quad L(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K\left(x_i \cdot x_j\right) \\ s.t \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \qquad 0 \leq \alpha_i \leq C \quad i = 1, 2, ...N \end{cases} \tag{4}$$

对非线性、不可分情况下的样本分类问题，SVM通过引入核函数及松弛变量

的方法进行解决. SVM最终的分类判别函数可表示为:

$$g(x)=\mathrm{sgn}\left(\sum_{i=1}^{sv}\alpha_i y_i K(x,x_i)+w_0\right) \tag{5}$$

其中, α_i 是w的非零元, sv为w中非零元的个数, $K(x,x_i)$ 为核函数。

核函数的具体形式对SVM的分类性能有较大影响, 但SVM的参数选择问题目前理论上尚未解决, 只能通过反复试验的方法选取. 我们比较了常用的线性核、多项式核、径向基(radial basis function, RBF)核及双曲正切核, 并最终选用RBF核函数:

$$K(x,y)=\exp(-\|x-y\|^2/2\sigma^2) \tag{6}$$

其中, σ=15. SVM中松弛变量的系数, 即惩罚因子C体现了SVM对训练集中样本的信任度, 并影响分类模型的推广能力, 本文选取C=2000。

D. 留一法

为检验选出的61个基因是否包含了完备的样本分类信息,本文以这61个基因作为样本的分类特征,采用支持向量机作为分类器进行样本的识别.然而由于基因表达谱数据集样本数量少,为了获得对分类错误率的可靠估计并同已有的研究成果进行比较,本文采用留一法进行检验,进行样本类型的识别,其原来和步骤如下所示:

在训练集上采用“留一法”(leave-one-out cross validation, LOOCV)进行样本识别,即在训练集上每次保留一个不同的样本作为测试样本,其余样本用做SVM的训练样本。重复该过程,直到训练集上所有样本均有一次机会被用做测试样本为止。记录所有被错误分类的样本数作为留一法的分类错误数,记为Err⊥LOOCV。

5.3.3 结论

最终我们得到的基因标签为:

表6 5个主要基因标签

编号	基因标签
1	X63629
2	H06524
3	H08393
4	R39209
5	M26383

我们从图上可以清楚看出,当取6个基因时,误差最小为6个,因此错判的概率为 $P=\frac{6}{62}=90.32\%$ 。

5.4 问题 3 模型与求解

5.4.1 理论分析

无论是肿瘤的分类与识别还是肿瘤的聚类分析,都是以大规模基因表达数据作为分析的基础。由于基因表达数据测定过程中需要经过多个步骤的操作,而每

一步都可能引入大量的噪声，这就使得基因表达谱数据属于强噪声数据，数据噪声的来源主要有如下几个方面：

(1)不同基因芯片闻的差异。

(2)样本制备手段的不同。

(3)基因探针与靶标基因的非特异性杂交。

(4)样本选择差异及基因组本身的不稳定性所造成的个体间的差异。

(5)样本组织中细胞组分的不同。

其中最主要的差异体现在个体基因组问由于个体多样性所造成的基因间的差别。同时由于样本获取和制各上的困难，因此目前的基因表达谱的样本数很少。这就造成了基因表达谱样本少，维数高，噪声大的特点。

5.4.2 模型建立

基因表达谱中不可避免地含有噪声（见 1999 年 Golub 在《Science》发表的文章），有的噪声强度甚至较大，对含有噪声的基因表达谱提取信息时会产生偏差。通过建立噪声模型，分析给定数据中的噪声能否对确定基因标签产生有利的影响。

在问题 2 中，我们已经建立了 SVM 分类器，对基因进行分类，然而这个模型显然是在没有噪声的情况下建立的，因此，在噪声下，该模型必须进行一定的修正。我们考虑，给定的数据中式存在噪声的数据，那么，引入 Δx 来表示对带噪声的实验数据的修正，x 仍表示带有噪声的实验数据，$\tilde{x}=x+\Delta x$ 则表示不带噪声的干净数据。

将无噪声的数据 $\tilde{x}=x+\Delta x$ 代入问题 2 中的模型中，则可以得到带噪声的 SVM 模型，即就是

$$\begin{cases} \max \quad \tilde{L}(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K\left(\left(x_i+\Delta x_i\right)\cdot\left(x_j+\Delta x_j\right)\right) \\ s.t \quad \sum_{i=1}^{N}\alpha_i y_i=0 \qquad 0\le\alpha_i\le C \quad i=1,2,...N \end{cases} \tag{7}$$

我们进一步对 $L(\alpha)$ 进行化简，那么可以得到

$$\begin{aligned} \tilde{L}(\alpha)&=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K\left(\left(x_i+\Delta x_i\right)\cdot\left(x_j+\Delta x_j\right)\right) \\ &=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \varphi\left(x_i+\Delta x_i,x_j+\Delta x_j\right) \\ &=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \varphi\left(x_i+\Delta x_i,x_j+\Delta x_j\right) \end{aligned} \tag{8}$$

假设噪声的成分很低，因此 $|\Delta x|$ 很小，则可将函数 φ 进行泰勒展开，得

$$\begin{aligned} \tilde{L}(\alpha)&=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \varphi\left(x_i+\Delta x_i,x_j+\Delta x_j\right) \\ &=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \left(\varphi\left(x_i,x_j\right)+\varphi_1^{'}\left(x_i,x_j\right)\Delta x_i+\varphi_2^{'}\left(x_i,x_j\right)\Delta x_j+O\Delta x^2\right) \\ &\approx L(\alpha)-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \left(\varphi_1^{'}\left(x_i,x_j\right)\Delta x_i+\varphi_2^{'}\left(x_i,x_j\right)\Delta x_j\right) \end{aligned} \tag{9}$$

由于函数 φ 是关于x轮转对称，并且考虑到 $\Delta x_i\approx\Delta x_j$，因此有

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(\varphi_1'\left(x_i,x_j\right)\Delta x_i\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(\varphi_2'\left(x_i,x_j\right)\Delta x_j\right) \tag{10}$$

从而，得到

$$\begin{aligned}\tilde{L}(\alpha)&\approx L(\alpha)-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(\varphi_1'\left(x_i,x_j\right)\Delta x_i+\varphi_2'\left(x_i,x_j\right)\Delta x_j\right)\\&\approx L(\alpha)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(\overrightarrow{\varphi_1'}\left(\overrightarrow{X}\right)\overrightarrow{\Delta X}\right)\end{aligned} \tag{11}$$

定义噪声矩阵为 $N(\alpha)=-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(\overrightarrow{\varphi_1'}\left(\overrightarrow{X}\right)\overrightarrow{\Delta X}\right)$，因此模型可以改写成

$$\begin{cases}\max\quad \tilde{L}(\alpha)=L(\alpha)+N(\alpha)\\ s.t\quad \sum_{i=1}^{N}\alpha_i y_i=0\qquad 0\le\alpha_i\le C\quad i=1,2,...N\end{cases} \tag{12}$$

此优化模型是带有参数 $\overrightarrow{\Delta X}$ 的优化模型，我们下面将从两个方面分析讨论：

为分析方便，不妨设 $\left|\overrightarrow{\Delta X}\right|\le\delta$，因此可以在 $[-\delta,\delta]$ 之间随机产生一组 $\overrightarrow{\Delta X}$，代入上述的优化模型，噪声的信息是很模糊的，并且噪声的定义也有很多，没有一个明确的定义，因此我们不妨认为SVM在无噪声的情况下能够找到全局最优的一组权重，从而确定分类的分解面，因此我们把影响SVM的权重的因素统统划分到噪声中，可能这种噪声与以往定义的噪声不同，但是它确实有一定含义的，分析这个噪声在理论上可以分类的准确度。

比较问题（2）中的优化模型和问题（3）中的优化模型可知，如果给定的数据是有噪声的，我们前面已经提出，在无噪声的情况下分类的准确度最高，那么问题（2）中代入的是含有噪声量的训练样本，而(12)所建立的模型中，$N(\alpha)=-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(\overrightarrow{\varphi_1'}\left(\overrightarrow{X}\right)\overrightarrow{\Delta X}\right)$，相当于对原来模型的一个修正，因此在理论上应该拥有最好的权重值。那么模型可以修正为

$$\begin{cases}\max\quad \tilde{L}(\chi)=L(\chi)+N(\chi)\\ s.t\quad \sum_{i=1}^{N}\chi_i y_i=0\qquad 0\le\alpha_i\le C\quad i=1,2,...N\end{cases} \tag{13}$$

这样可以通过求解非线性优化模型，得到一组最优解为 $\chi^*=[\alpha^*,\overrightarrow{\Delta X^*}]$，那么所得到的 α^* 为找到的最优权值，我们定义 $\overrightarrow{\Delta X^*}$ 为可调控因子，为此我们大胆设想，如果我们把噪声所引起的坐标漂移的范围设成沿 $\overrightarrow{\Delta X^*}$ 的方向，那么噪声引起的坐标漂移就为 $k*\overrightarrow{\Delta X^*}$。

5.4.3 实验仿真

从实验仿真中，我们发现 SVM 对线形函数的效果并不是很好，在无噪声输入时，“留一法”的误差有时能达到 27，这个使得效果大大减弱。但是我们通过大量的实验仿真发现，噪声修正的输入似乎可以降低噪声。然而并不是随便加入噪声就可以大幅度减少出错量。我们给出 $\left|\overrightarrow{\Delta X}\right|<\delta$ 中 δ 的范围，似乎有规律可循，将实验的仿真结果列入下表

表 7 噪声模型的比较

δ 的取值	出错情况	描述
0	25	无噪声的模型

0.001	16	有噪声模型
0.005	11	
0.01	10	
0.1	11	
0.5	12	
1	16	

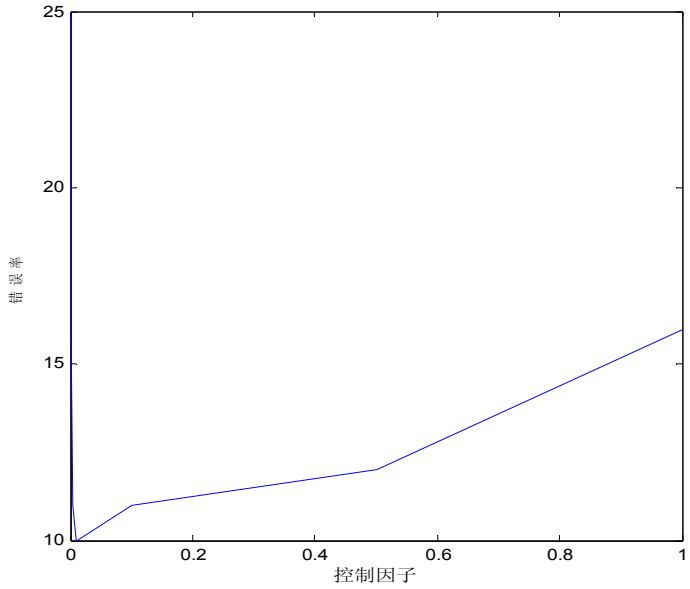


图 6　控制因子与误判数关系图

将其横坐标对数化，得到

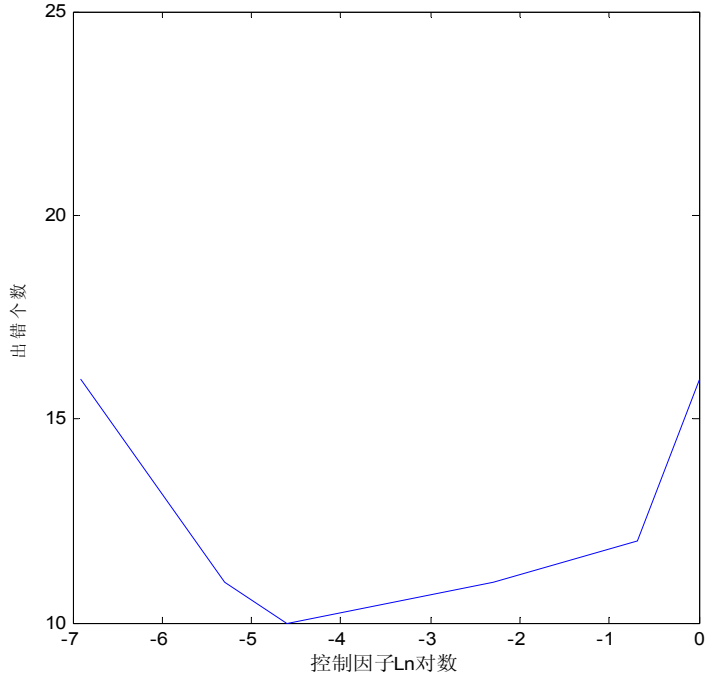


图 7　控制因子与误判数的对数关系图

从图中可以看出，随着控制因子增加误差并不是单调递减，而是出现先减少后增加的趋势，这说明总能找控制因子的一个合理范围能够使得误差达到最小，可惜的是，不同的样本的控制因子范围不同，也没有一个很好的解析表达能够很好地刻画，但是可以通过大量实验仿真得到控制因子的答题范围。另外一个就是方向的问题，$\overrightarrow{\Delta X}$ 是一个向量，即使选在 δ 在合理的范围之内，其实有时候也未

必能够保证得到最少的出错数。本文给出一个方向的判别标准，然而，通过仿真发现这个标准效果很好，但是遗憾的是，由于时间有限，并没有从理论上加以证明。

定义$\overrightarrow{\Delta X}$的方向应该和$\overrightarrow{\Delta\zeta}$的方向相反。其中

$$\overrightarrow{\Delta\zeta} = mean(\zeta) - mean(\{\zeta - i\}) \tag{14}$$

这个公式表示，$mean(\zeta)$样本的均值，$mean(\{\zeta - i\})$表示去掉该样本后，剩余样本的均值。这其实就是表明了，正常情况下，样本应该大体分布在均值附近，而偏移均值很远就会造成很大的$\overrightarrow{\Delta\zeta}$，可以认为是一个盲点，即可以认为是数据误差或者是数据噪声引起，此时应该将这个因素添加到优化模型中来，对原来的模型进行修正，即产生了上述的$N(\alpha)$。

5.4.4 结论

上面我们所做的仿真，得到了两个结论：（1）某个$k*\overline{\Delta X^*}$总是比随机噪声的效果好，即SVM的分类误差会减小；（2）总会存在比较理想的k的范围，使得$k*\overline{\Delta X^*}$能够比不使用噪声模型或者随机噪声模型的误差要小。

5.5 问题 4 模型与求解

5.5.1 理论分析

本题的解答是通过建立一个数学模型，根据若干基因标签来帮助临床诊断结肠癌。通过前面的分析已经确定了三个基因标签，下面阐述如何通过测试样本在这三个基因上的表达水平来决定该样本是属于正常样本还是结肠癌样本。

5.5.2 模型建立

采用递归划分的思想来建立该诊断模型，根据前面确定的基因标签确定一颗二叉树，图 8 为一个典型的二叉树模型，可以用于基因诊断，其中每个基因标签，作为一个内部分类节点，叶节点表分类结果。

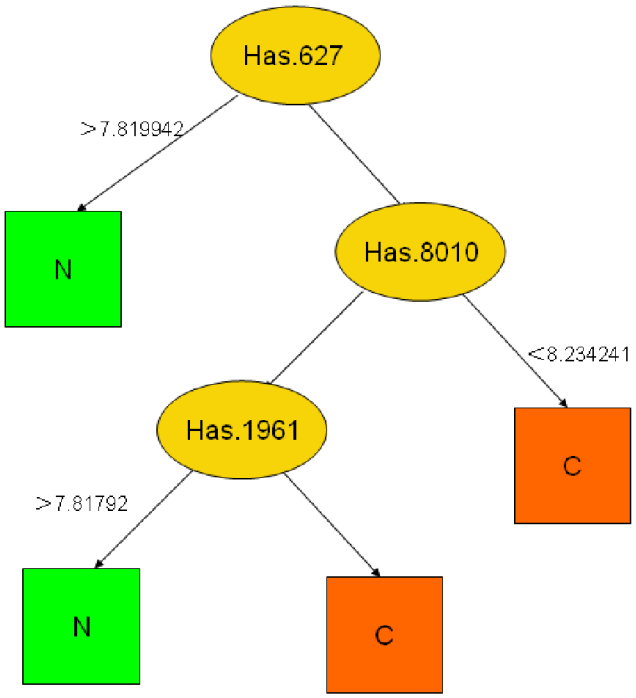
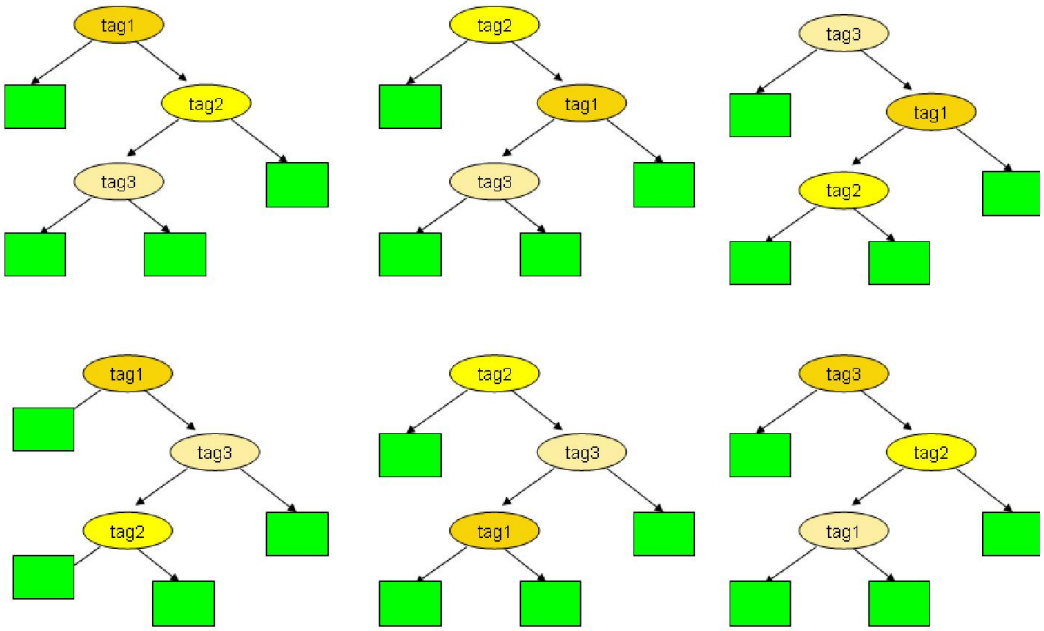


图 8　基于二叉决策树的基因诊断模型

样本从根节点开始往下移动，直到叶节点才停止。在每个内部节点，需要做一次分类判断，分类的依据是该样本在该节点（基因标签）上的表达水平。这里每个内部节点需要建立一个分类阈值以及阈值分类方向（即样本表达水高于阈值时，向左子树移动还是向右子树移动）用以区分样本的类别。这样，模型的参数包括如下三个内容：

1. 内节点的顺序问题，需要确定二叉树的结构。包含以 tag1，tag2，tag3 为内节点的二叉树有以下几种结构。



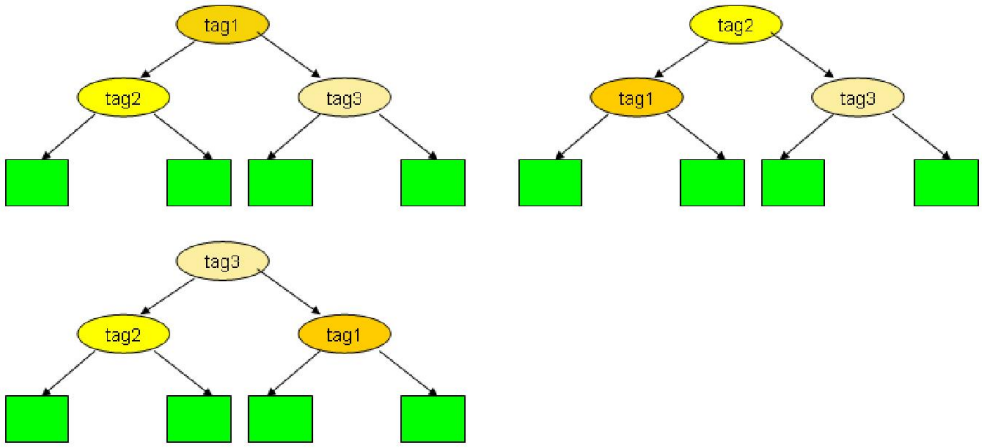


图 9　三个内结点组成的所有可能的二叉决策树

2. 阈值分类方向问题，要确定是利用样本在节点（基因）上的表达水平大于阈值这一条件还是利用表达水平小于阈值这一条件，以及大于（或小于）节点阈值时应该向左子树移动还是向右子树移动的问题。
3. 每个节点的阈值的确定，如何选取一个较好的阈值以提高诊断的可靠性。

参数的确定：

1. 内节点的顺序问题，根据前面的题一、题二，已经得到了一组基因标签，并且确定了基因标签的特征表现的排序，假如基因标签 tag1 特征表现最强，则说明 tag1 在所有基因中区分正常样本和癌症样本的能力最强。这里可根据得到的基因标签的特征表现强度来建立二叉树。如果一个基因的特征表现强度最大，则它对样本分类的贡献能力最强，那么可以将它放到根节点。同理按照基因标签的表现能力的衰减可以逐级往下排列节点。例：假如确定基因标签的表现能力按大到小的排列顺序为 tag1>tag2>tag3 ，那么就建立如下图所示的二叉树。

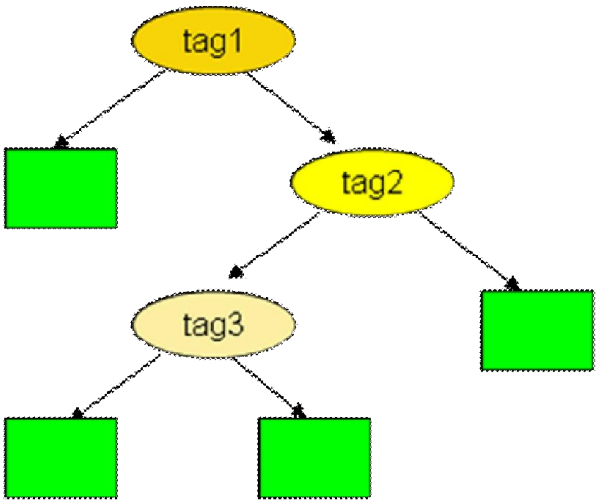


图 10　最优二叉决策树

2. 阈值的选取，为了确定阈值，可以将样本分为训练集和测试集，训练集用于模型参阈值的确定，测试集用于检验模型的可靠性。取 15 个正常的样本的特征表现水平向量 X1[1:15]和 30 个结肠癌样本的特征表现水平向量 X2[1:30]作为训练集(这里的 X1[n] ,X2[m],分别为训练集和测试集在基因

标签 tag1、tag2，tag3 的表现水平向量)，则训练集集合为 X=[X1,X2]，采用支持向量机对训练集 X 进行训练，可以得到一个对训练集 X 的一个最优划分。然后，更新训练集在每个基因标签上的表现水平，具体跟新方法是针对每一个基因标签（如 tag1）设置该基因在所有的训练集上的表现水平为变量 x，令 x 从该基因原先训练集中最低表现水平(Etag1_min)到最大表现水平(Etag_max)变化（取步长△x）对每个 x 值，将更新后的训练集带入前面的支持向量机进行检测，并记录出错的样本个数。这样总可以找到一个 x=Etag1_threshold 使得出错的样本个数最少。取 tag1 结点的阈值为 Etag1_threshold,同理可得 tag2、tag3 结点的阈值。算法流程如图 11 所示：

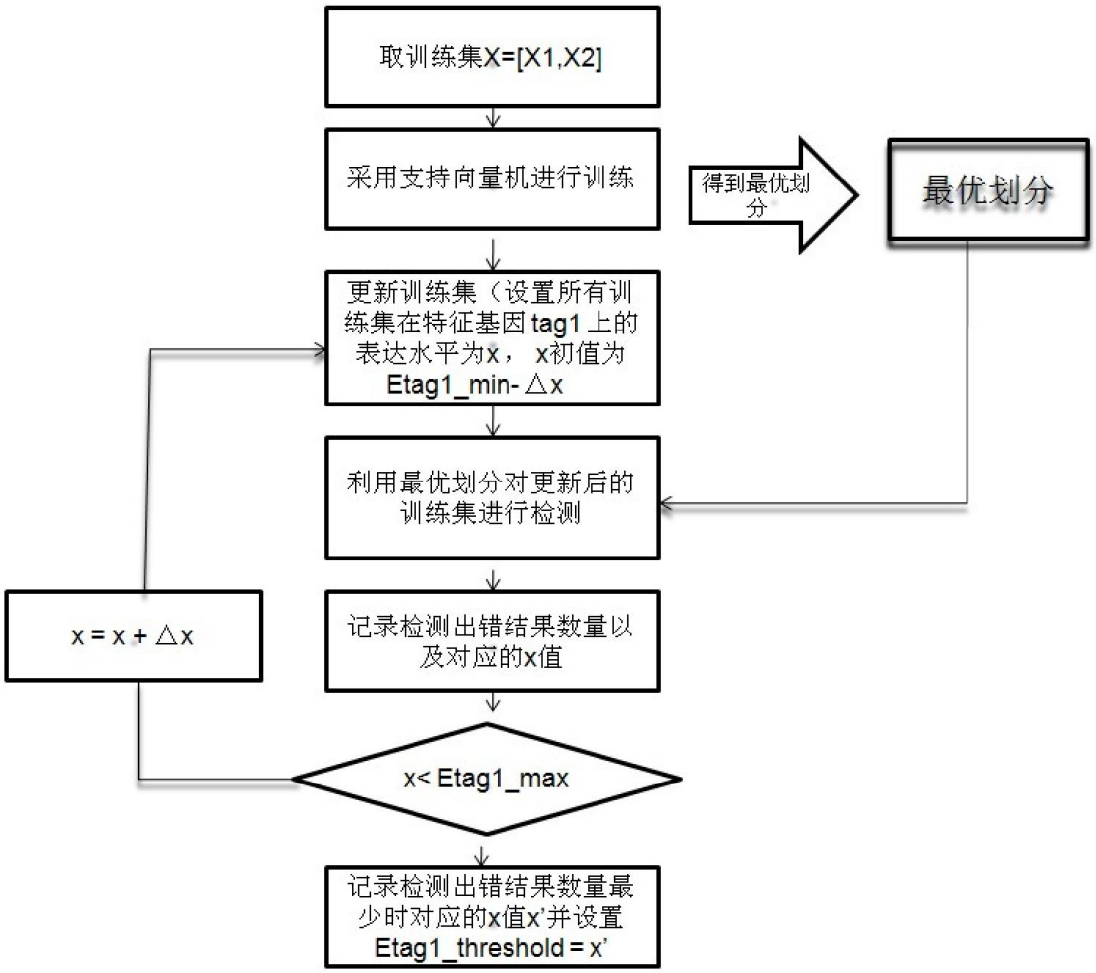


图 11　　阈值选取算法流程（以 tag1 为例）

3. 阈值分类方向问题，为了方便讨论，规定类别为正常样本的分类结果总是偏向当前结点左子数。考虑一个测试集样本 TS，设其在 tag1 上的表达水平为 TS_Etag1，在经过结点 tag1（阈值为 Etag1_threshold）时，有四种选择方案，分别是:

   A. TS_Etag1>Etag1_threshold, TS 进入左子树。
   B. TS_Etag1>Etag1_threshold, TS 进入右子树。
   C. TS_Etag1<Etag1_threshold, TS 进入左子树。
   D. TS_Etag1<Etag1_threshold, TS 进入右子树。

为了确定每个节点上的路径选择方案，需要利用到训练集。对于每个基因标签，计算它在训练集中的正常样本中表达水平大于和小于阈值的概率，记为 *Pnb*，

Pns=1-Pnb,在计算它在训练集中的结肠癌样本中表达水平大于和小于阈值的概率，记为 *Pcb*，*Pcs=1-Pcb*。根据信息论理论该基因标签在正常样本中表达的不确定性可以用如下公式来衡量：

$$I_n = P_{nb}\log(1/P_{nb}) + P_{ns}\log(1/P_{ns}) \tag{15}$$

同理该基因在结肠癌样本中的表达的不确定性为：

$$I_c = P_{cb}\log(1/P_{cb}) + P_{cs}\log(1/P_{cs}) \tag{16}$$

所以当 $I_n < I_c$ 时说明该基因标签在正常样本中表达的不确定性较小，那么可以向确定为正常样本分类的方向选择（及选择向左子树方向选择），那么就在方案 A、C 中进一步选取，反之，如果 $I_n \geq I_c$，这需要在方案 B、D 中进一步选取。先假设 $I_n < I_c$，那么如何确定是选放案 A 还是方案 C 呢？*考虑 Pnb，如果 Pnb>1/2* 说明该基因标签在所有正常样本中大于节点阈值的概率大，故选择方案 A，反之选择方案 C。对于 $I_n \geq I_c$ 的情况，道理一样。下面列出了建立节点路径选择方案的方案：

a. $I_n < I_c$，并且 *Pnb>1/2* 选择方案 *A*

b. $I_n \geq I_c$，并且 *Pcb>1/2* 选择方案 *B*

c. $I_n < I_c$，并且 *Pnb<=1/2* 选择方案 *C*

d. $I_n \geq I_c$，并且 *Pcb<=1/2* 选择方案 *D*

根据上面的方法，可以建立类似下图的分类树，其中白框和黑框代表两种不同的类别（在本题中就是正常样本和结肠癌样本）。

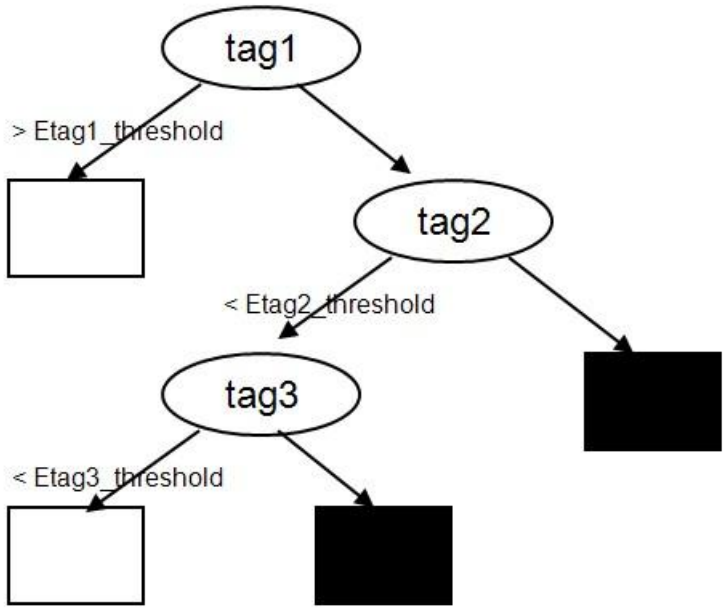


图 12 带阈值的二叉决策树

到这里，一个基因诊断模型已经建立好，可以利用它来进行基因诊断了，对于一个给定的测试样本 TS，诊断方法如下：

从根结点开始，比较 TS 在基因标签 tag1 上的表达水平 TS_Etag1，如果 TS_Etag1> Etag1_threshold,则进入左子树,为叶节点，停止比较。

否则进入右子树 tag2，同 tag1 一样进一步进行判定，直到叶结点。

到了叶节点，根据叶节点所属类别，判定 TS 的类别，如果叶节点属于正常类别，则判定 TS 为正常样本，否者 TS 为结肠癌样本。

根据前面的方法建立本题中数据相对应的诊断模型。

首先，选取训练集,这里抽取 excel 数据表中的 54 个样本作为训练集，这 54 个 sample 如下：

normal1 normal2 normal3 normal4 normal5 normal6
normal7 normal8 normal9 normal10 normal13 normal14
normal15 normal16 normal17 normal18 normal19 normal20
normal22
cancer1 cancer2 cancer3 cancer4 cancer5 cancer6
cancer7 cancer8 cancer9 cancer10 cancer11 cancer12
cancer13 cancer14 cancer15 cancer16 cancer17 cancer18
cancer19 cancer20 cancer23 cancer24 cancer25 cancer26
cancer27 cancer28 cancer29 cancer30 cancer31 cancer32
cancer36 cancer37 cancer38 cancer39 cancer40

剩余的 8 个样本作为测试集。由题二可知，特征基因分别为 M26383，R39209，H08393，H06524，X63629（按特征表现由强到弱排列）。然后，根据前面的阈值选定算法确定这五种基因结点的阈值，利用 matlab 程序（见附件 matlab4.rar）实现上述算法，计算得：

表 10 基因标签与阈值表

编号	基因标签	阈值
1	M26383	8.7859
2	R39209	4.5860
3	H08393	6.5035
4	H06524	5.9461
5	X63629	8.1686
注：这里直接使用 excel 表格中提供的原始数据进行计算		

接着确定各个结点的阈值方向，以第一个结点的阈值方向确定为例，上述 54 个训练集样本在基因 M26383 上的表现水平如下：

表 11 训练集的基因在 M26383 上的表现水平

样本	在基因M26383上的表现水平	样本	在基因M26383上的表现水平
normal1	4.56924803	Cancer9	7.209855412
normal2	5.168270966	Cancer10	7.798042412
normal3	5.086295483	Cancer11	7.720423836
normal4	5.830832074	Cancer12	10.5778123
normal5	4.846367946	Cancer13	6.250298418

normal6	5.495055528	Cancer14	7.402052749
normal7	3.507953169	Cancer15	7.738953565
normal8	4.807805694	Cancer16	8.531537419
normal9	5.313155136	Cancer17	6.844533164
normal10	4.986297154	Cancer18	8.350613268
normal13	5.204277685	Cancer19	8.282798859
normal14	5.334094474	Cancer20	7.541309615
normal15	6.266552318	Cancer23	8.150489865
normal16	5.786596362	Cancer24	10.20324864
normal17	7.898367091	Cancer25	6.244387622
normal18	6.613531653	Cancer26	9.97660164
normal19	6.831798088	Cancer27	7.83103008
normal20	9.335376397	Cancer28	6.335189346
normal22	9.147538843	Cancer29	10.82452619
Cancer1	6.24697806	Cancer30	7.410398349
Cancer2	8.258842769	Cancer31	9.526880647
Cancer3	8.612334067	Cancer32	8.077863842
Cancer4	11.02021778	Cancer36	7.960878002
Cancer5	9.228091148	Cancer37	6.872644102
Cancer6	7.455748459	Cancer38	6.711236768
Cancer7	6.915520901	Cancer39	6.964846228
Cancer8	6.544694023	Cancer40	9.019816566

前 19 个是正常样本的 M26383 表现水平，后 35 个是结肠癌样本的 M26383 表现水平，则：

Pnb = 2/19=0.8947　Pns=1-Pnb = 0.1053

Pcb = 8/35=0.2286，Pcs=1-Pcb =0.7714 。

$$I_n = P_{nb}\log(1/P_{nb}) + P_{ns}\log(1/P_{ns}) = 0.4856$$

$$I_c = P_{cb}\log(1/P_{cb}) + P_{cs}\log(1/P_{cs}) = 0.7756$$

满足 $I_n < I_c$，又因为 *Pnb= 0.8947 >0.5* 所以选择方案 *A*。

及当样本在 M26383 上的表现水平>8.7859 时应该选择进入左子树，否则进入右子树，经过 M26383 阈值的划分，左子树包括了训练集中的 17 个正常样本和 27 个结肠癌样本，右子树中包含了 2 个正常样本和 8 个结肠癌样本。需要在子结点中继续划分。接着取 R39209 基因对 M26383 的左右子结点进一步划分（这里由于左右结点都包含两类结点，所以 R39209 要在 M26383 的左右子结点中都要做划分，即 R39209 需要用两次）。重复前面步骤知道所有训练集被区分开并给予归类

5.5.3 结论

最后得到如下的二叉树：

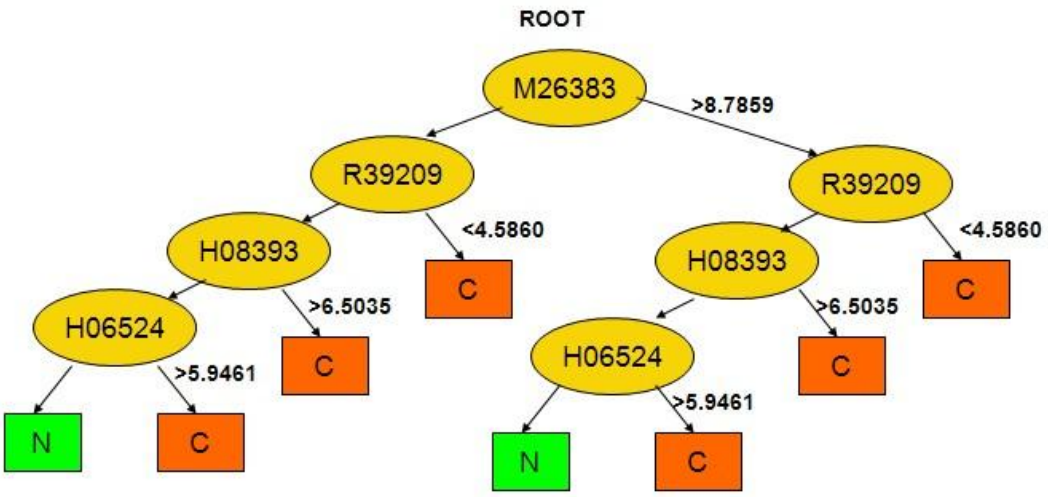


图 13　　结肠癌基因诊断的二叉树树模型

观察上图可知，基因 X63629 没有被用上，实际上在二叉树模型建立时到第四层结点时已经可以正确的将 54 个样本正确归类了，不需要基因 X63629 做进一步判断了。这种现象可能和样本选取有关，在前面求解五个特征基因时选取的是所有的样本都用上了，而这里建立诊断模型时只用了其中的 54 个样本。

最后，利用上面建立好的诊断模型对测试集样本进逐个测试，例如取样本 normal11 它在 M26383，R39209，H08393，H06524 上的表现水平分别是：4.934457528 ，8.271074832， 6.204179852， 9.239464277，6.735606792，进行诊断时先 4.934457528 和 M26383 结点阈值 8.7859 比较，4.934457528<8.7859，则进入左子树 R39209 ，又因为 8.271074832>4.5860 所以进入左子树 H08393，又因为 6.204179852<6.5035，所以进入 H08393 的左子树 H06524，在结点 H06524 处 9.239464277>5.9461，则最后进入左节点（叶节点）并判断为正常基因（结果正确），按同样的方法对其他的测试样本进行检查，比价检查结果是否正确，最后发现 8 个测试样本中只有样本 normal12 检查出错。这样初步估计诊断的可靠性为 7/8 = 87.5%。

该基因诊断模型的采用递归划分的思想对样本进行划分，同时具有统计意义， 模型本身比较直观。本文中的阈值选取是利用支持向量机来实现的，由于计算叫复杂，为了减少计算时间，在求解阈值的时候 x 的步长△x 没有取得很小（本文中△x=0.01 ,理论上要满足原始数据的精度），因此，对判断的准确性有一定的影响，如果减小△x 会进一步提高模型的可靠性。总的说来，本模型对提供的样本数据能做出较为可靠的判断，当然，还有改进的地方，如基因标签的选择，采取基因表征能力排序方法等还是值得研究的。

6 附录

以下是所有数据经 Relief 算法排序后的基因顺序。

编号	EST name	GenBank Acc No
1	Hsa.37937	R87126
2	Hsa.627	M26383
3	Hsa.1832	R27369

4	Hsa.36952	H43887
5	Hsa.692	M76378
6	Hsa.8147	M63391
7	Hsa.6814	H08393
8	Hsa.1131	T92451
9	Hsa.3305	X12369
10	i	
11	Hsa.36689	Z50753
12	Hsa.2344	X86693
13	Hsa.2291	H06524
14	Hsa.2097	M36634
15	Hsa.404	L07648
16	Hsa.601	J05032
17	Hsa.41260	L11706
18	Hsa.549	R36977
19	Hsa.8125	T71025
20	Hsa.2456	U25138
21	Hsa.6596	R76254
22	Hsa.27721	R50505
23	Hsa.4689	T95018
24	Hsa.2644	X54941
25	Hsa.24506	R44418
26	Hsa.831	M22382
27	Hsa.2471	H20512
28	Hsa.1047	R84411
29	Hsa.2645	X54942
30	Hsa.3349	X15882
31	Hsa.27686	H20426
32	Hsa.773	H40095
33	Hsa.1221	T60155
34	Hsa.33	U05040
35	Hsa.11616	T60778
36	Hsa.2451	U22055
37	Hsa.28939	R60877
38	Hsa.1130	Z24727
39	Hsa.957	M26697
40	Hsa.2928	X63629
41	Hsa.1013	T61661
42	Hsa.1073	X12466
43	Hsa.832	T51023
44	Hsa.1902	L05144
45	Hsa.466	U19969
46	Hsa.1985	T52185

47	Hsa.2715	H77597
48	Hsa.558	R34698
49	Hsa.2372	D16469
50	Hsa.662	T86749
51	Hsa.579	M80815
52	Hsa.2250	U17899
53	Hsa.3016	T47377
54	Hsa.41338	D31716
55	Hsa.1205	R08183
56	Hsa.821	X14958
57	Hsa.3306	X12671
58	Hsa.72	D29808
59	Hsa.6317	R39209
60	Hsa.41280	Z49269
61	Hsa.25322	R44301

7 参考文献

[1] 王树林等，肿瘤信息基因启发式宽度优先搜索算法研究，计算机学报，31（4）：636-649，2008.

[2] 阮晓钢等，基于基因表达谱的肿瘤特异基因表达模式研究，中国科学 C 辑 生命科学，36 (1): 86~96，2006.

[3] Golub T R，Slonim D K，Tamayo P，Huard C，Gaasenbeek M，Mesirov J P.Coller H，Loh M L，Downing J R，Caligiuri M A，BIoomfield C D，Lander E S．Molecular elassification of cancer：Class discovery and class prediction by gene expression monitoring．Science，1999，286l 531-537.

[4] 李颖新，阮晓钢．基于支持向量机的肿瘤分类特征基因选取．计算机研究与发展，2005，42(i0)：1796—1801).

[5] 边肇祺，张学工．模式识别(第2版)．北京：清华大学出版社，2000：180-183

[6] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Machine Learning, 2000, 46(13): 389~422[DOI].

[7] Vapnik V N. Statistical Learning Theory. New York: Wiley Inter-science, 1998.

[8] Heping Zhang, etc. Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci USA.98(12): 6730-6735. 2001.