全国第七届研究生数学建模竞赛

题 目 基于临床与基因图谱的结肠癌基因标签提取

摘 要

由于基因间的调控和相互作用表现为“功能基因组合”形式，基因的功能与作用是集体作用的结果，而非单个基因单独作用的结果，表现在分类特征对样本的分类能力方面就是以特征集合的形式整体体现出来的。根据这个生物学知识，本文考察由多个基因构成的基因簇作为区分正常人和癌症患者的分类因素，利用独立成分分析(ICA)技术对已给出的基因表达采样数据进行分析，最大程度地降低基因之间强烈的相互影响，从而获得对判断是否患有肿瘤或者癌症的最有直接关系但数目较少的潜在因素，即基因簇信息。随后，我们采用了支持向量机(SVM)依据提取出的潜在因素（基因簇）进行分类，筛选出致病的癌症基因15个。另外，我们还运用基于灵敏度的支持向量机对基因本身进行分类，而不是基于基因簇。利用得到的结果与基于独立成分分析的方法所提取的基因提供比较。发现所筛选的基因簇中有三个基因与灵敏度支持向量机方法筛选的基因相同。

对预处理过后的1908个基因，通过独立成分分析提取出61个基因簇，这些基因簇中含有与分类无关的基因簇，即噪声，以及与分类相关的分类因素5个。事实上，为了能够得到最好的分类因素，我们将问题转化为一类信号稀疏表示的优化问题。此外，为了进一步进行基因分类，我们利用含噪声的ICA和带松弛因子的非光滑优化模型研究带有噪声的基因图谱信息。通过含噪声模型与不含噪声模型进行对比，说明含噪模型的优势。

最后，借助于条件概率模型，对病人数据进行了筛选，将临床结论与基因图谱相结合，通过已有文献以及生物信息网站所获取资料发现，所筛选的大部分基因标签与当今临床医学所得到的直肠癌研究结论相吻合。

关键词：含噪 基因簇 独立成分分析 支持向量机 非光滑优化模型 临床 基因标签

参赛队号 10291006

队员姓名 邵伟 孟秋池 葛成伟

参赛密码 _____

（由组委会填写）

中山大学承办

一、问题的重述

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。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。通常由于基因数目很大，所以在判断肿瘤基因标签的过程中，我们需要剔除掉大量“无关基因”，从而大大缩小需要搜索的致癌基因范围。

从 project_data.txt 数据中获取的基因表达谱中的数据中包含 62 个样本(其中 22 个为正常人样本，40 个人为癌症病人样本)，每个样本中包含 2000 条基因数据，我们着重需要解决以下几个问题：

(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。对于给定的数据需要选择最好的分类因素；

(2) 相对于基因数目，样本往往很小，对于给定的结肠癌数据需要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，对含有噪声的基因表达谱提取信息时会产生偏差，需要建立噪声模型去分析给定数据中的噪声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（如 APC、RAS 基因）与某种癌症的关系密切，需要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。

二、模型假设

1. 基因表达谱数据中虽然含有噪声，但随机噪声的强度不会淹没真正的基因信息。

2. 不同基因信号之间的冗余关系是存在的。

三、数据预处理

基因芯片经激光扫描仪扫描，再经图像分析软件进行处理，得到反映基因表达水平的数据序列。这些数据用于差异表达基因的鉴别和基因表达模式的分析之前，还需要进行初步的处理。如为了从生物学角度上更好地解释及使数据满足特定的数据分布，需要对荧光强度数据进行对数转换；实验中系统误差的存在使得不能对不同样本的数据进行直接比较，因此针对系统偏倚产生的原因而进行数据归一化是必要的，也是数据预处理中重要的一个步骤。

（一）数据的对数转换

对数据进行对数转换是基于以下一些方面的原因。

首先是在生物学上易于理解和解释。假设两个基因在对照样品中的背景校正

强度值均为 1000，而在另外一种实验条件下的强度值分别为 100 和 10000。如果从对照与实验的绝对值来看，一个基因表达的变化远远大于另一基因，即 10000-1000>>1000-100。但是，从生物学的角度出发，两个基因变化的是相等的，都是 10 倍的变化。用对数转换可以消除这种由两个相对变化间的不成比例所引起的误导。例如，对数据进行以 10 为底的对数变换，则

$$\lg 100 = 2$$

$$\lg 1000 = 3$$

$$\lg 10000 = 4$$

可以看出，基因的变化是相等的，只是方向不同，一个增大，另一个减小。对数变换减弱了数据的平均值和方差，使得表达的变化独立于其产生的强度位置，从而使得低强度值与高强度值发生的倍数变化具有可比性。

另外，对数变换使得数据的分布具有对称性和接近正态分布性质，而一些常用的统计方法，如 t 检验、F 检验等方法都要求数据满足正态分布或近似正态分布。由于本问题中所提供的数据已经是对数形式的，所以可以忽略这一步。

（二） 重复数据的合并

重复的测量可以用于估计实验中的噪声，比较不同处理组间和处理组内的变异。然而，在特定的条件下把所有的重复值合并成一个数值可能更为方便，而这一个值就是给定基因（条件）的代表。根据不同的情况，这些重复测量可能是同一芯片上的重复点，或是同一基因在不同芯片上的测量值。通常的合并是指计算这些重复值的集中趋势指标，如均数、中位数或众数。

（三）数据归一化

系统误差使得采集到的数据可能含有奇异样本数据，所谓奇异样本数据指的是相对于其他输入样本特别大或特别小的样本矢量。奇异样本数据的存在会影响特征基因的提取。所以，在数据预处理部分，需要对原始数据进行归一化。归一化的具体作用就是归纳统一样本的统计分布性。归一化在 0~1 之间是统计的概率分布，归一化在-1~+1 之间是统计的坐标分布。例如规整原数据到$[0,1]$内，这样可以降低奇异样本数据对整体的误差影响，从而更加有效地提取特征基因。另外，数据归一化对于独立分量分析（ICA）、支持向量机（SVM）数据处理也是有帮助的。

首先，根据附件的文件说明，我们需要对 project_data.txt 里的数据进行以下预处理：

1. 在project_data.txt数据文件中，第二列为UMGAP，HSAC07 或者i的数据是和RNA控制相关的，对下面所做的工作没有关系，为冗余数据，所以需要把这些数据去除。

2. 基因芯片探针探测到的序列表明了基因的表达水平，有些数据可能是同一基因探针的重复点，也有可能是同一基因在不同基因探针上的探测值。因而，对于project_data.txt中基因相同的序列，需消除重复表示，采用了类均值算法，对其进行取平均或取中值处理，给出特定基因的唯一表达数据。

以上两步的数据预处理可以保证：处理后的数据较真实地反映了不同基因的不同表达水平。

通过以上的预处理，原基因数据从2000个基因减少到了1908个。实验表明，1908个基因数据为可靠性较高的数据。

其次，进行数据归一化处理。采用的归一化映射为：

$$f: x \to y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.1}$$

公式(3.1)当中，$\vec{x}, \vec{y} \in R^d$，$x_{\min} = \min(x)$，$x_{\max} = \max(x)$，x，y 分别是 $\vec{x}, \vec{y}$ 相对应的坐标分量。归一化的效果是原始数据被规整到[0,1]内。

四、基于 ICA 技术确定分类因素

我们知道在癌症形成的过程中，蛋白质能够改变细胞之间的信号，这种信号传递功能通常称为细胞内信号传导。这种信号传导在癌症发生过程中很关键，因为肿瘤细胞的异常生长大多由细胞内信号传导分子的异常功能形成。此外，一些变异基因产生的酶（一种蛋白质）会导致正常基因的错误转录和表达，从而生成导致癌症形成的蛋白质，而这种现象会使得探测到的正常基因有异常的表达。

由此可见，蛋白质在癌症的发生中起到重要作用，所以从研究致癌的蛋白质或者酶的角度出发去进行分类会更加符合生物医学的事实，所得到的与肿瘤或者癌症直接相关的突变基因会更加准确，避免了直接从统计意义上的分类角度出发而导致的在突变基因筛选的错误。

虽然有人尝试评估出人类基因的总数(略多于22000个)，但是想要从中推测出人类基因组成所编码的不同蛋白质的总数仍然十分困难，最简单的评估方法是假设每个基因都仅仅编码一个蛋白质，但是这种编码模式过于简单，它忽略了蛋白质形成的复杂性，而且每个基因转录得到pre-mRNA可能具有几种选择性的拼接模式。事实上，每个蛋白质是由若干个基因组合（称为一个基因组或者基因簇）来决定并形成的，从基因簇的角度进行研究必定会更加真实而有效。

显然，基因图谱中的基因之间存在着很强的关联，为了能够得到更好地得到分类的因素，我们需要发掘基因数据之中的潜在因素。依据上述思想，本文将基因簇（基因组）作为区分正常人和癌症患者效果的分类因素，利用独立成分分析(ICA)技术对已给出的基因表达采样数据进行分析，最大程度上降低基因之间强烈的相互影响，从而获得对判断是否患有肿瘤或者癌症的最有直接关系但数目最少的潜在因素，即基因簇信息。

ICA 是近些年来发展起来的一种统计方法，其基本思想是用一组基函数来表示一系列随机变量，并且假设基函数彼此之间是统计独立的或者尽可能独立的。独立成分分析已经成为近年来神经网络、高级统计学、医学和信号处理等研究领域中最令人振奋的主题之一，科学家们在不断努力地从事这个方法的研究与探索工作，并取得了众多的研究成果。

ICA 最简单的形式是：给定 m 个可观察基因表达信号 $\bar{x}_1, \bar{x}_2, ..., \bar{x}_m$，假设它们为 n 组未知的相互独立的基因簇信号 $\bar{s}_1, \bar{s}_2, ..., \bar{s}_n$ 的线性组合，这些独立成分相互之间满足统计独立的假设，且都具有零均值。那么，将基因表达信号和基因簇用矩阵的形式进行表示为：$X = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_m)^T$，$S = (\bar{s}_1, \bar{s}_2, ..., \bar{s}_n)^T$，则 S 和 X 之间的关系如下：

$$X = AS \tag{4.1}$$

其中矩阵 A 称为混合矩阵。ICA方法就是在混合矩阵 A 和独立成分矩阵 S 未知的情况下，根据观测数据矩阵 X 确定分离矩阵 W，使得：

$$\hat{S} = A^+ X = WX \tag{4.2}$$

矩阵 $\hat{S}$ 是对 S 的最优估计。其中 $\hat{S} = (\bar{\hat{s}}_1, \bar{\hat{s}}_2, ... \bar{\hat{s}}_n)^T$ 且：

$$\hat{S}_{i,t}=\sum_{j=1}^{m}W_{i,j}*X_{j,t} \tag{4.3}$$

其中，$t=1,2,\ldots,sn$，sn为采样数。从(4.3)式可以看出各独立的基因簇信号是可以通过对所有的基因表达信号进行线性组合的方式获取。这里我们采用FASTICA算法计算上式中的W,可以获取隐藏在基因信号中的多组基因簇。

FastICA算法是一种快速寻优迭代算法，可以认为是梯度算法的一个在计算上优化了的版本，与普通的神经网络算法不同的是这种算法采用了批处理的方式，即在每一步迭代中有大量的样本数据参与运算，它不但能够在在线学习而且收敛快速。FastICA算法有基于四阶累积量、基于似然最大、基于负熵最大等众多形式，这里我们介绍下基于负熵最大的FastICA算法，即信息传输极大原则去计算独立成分S。其具体实现步骤如下：

1）给定初始值W_0(可以是随机的)，输入观测矩阵X；

2）由观测信号估计出相关矩阵R_x，并对R_x进行QR分解得到白化矩阵P；

3）用白化矩阵P对$\vec{X}$进行白化处理；

4）计算

$$Y=WX,\ \eta,\quad \Delta W=\eta(1-\varphi(y))$$

其中$\varphi(y)=[-\frac{g_1^{''}(y_1)}{g_1^{'}(y_1)},\ldots,-\frac{g_n^{''}(y_n)}{g_n^{'}(y_n)}]^T$称之为评价函数，我们将对应的非线性函数选为sigmoid，即

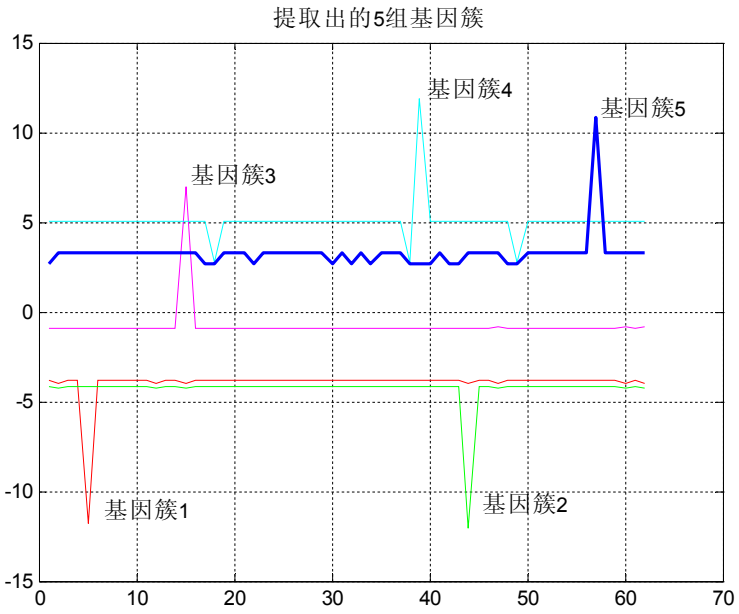
$$g_i(y_i)=(1+e^{-y_i})^{-1}$$

5）计算$\Delta W=\eta(1-\varphi(y)y^T)W$；

6）计算$W=W+\Delta W$，重复4,5,6步直到收敛。

运用上述算法，对经过数据预处理后的数据进行独立成分分析后，计算得到隐藏在基因表达数据中的相互独立的基因簇共有61组，根据后面的分类分析，其中对分类关联最大的5组基因簇的数据如下图(1)所示。我们提取基因簇1中的数据有62维如下所示(其它的60组略去，不显示在本文中)：

$ICA1$= [-3.8391 -3.9836 -3.8391 -3.8391 -11.7857 -3.8391
-3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.9836 -3.8391
-3.8391 -3.9836 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391
-3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391
-3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391
-3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391
-3.8391 -3.8391 -3.9836 -3.8391 -3.8391 -3.9836 -3.8391
-3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391 -3.8391
-3.8391 -3.8391 -3.8391 -3.8391 -3.9836 -3.8391 -3.9836];



图（1）提取出的 5 组基因簇的数据

从图(1)中可以明显发现，在同一个基因簇上，大部分的数据值是相似的，只有少部分元素具有较大的变化，这说明个别基因对整个基因簇有较大的影响度，很可能就关系到某些与致癌显著有关的基因，也就是致癌基因“标签”。

以基因簇为分类因素出发，这样就确定了与致癌直接相关的基因簇的范围，最大程度上建立了基因之间的相互关系。这样做大大缩小了寻找致癌相关基因的搜索范围，并且从更深层次上挖掘基因之间的复杂关系，可以得到与某种癌症相关联的因素，所得结果必将更加符合生物医学理论与实际。

五、基于灵敏度的 SVM 的基因标签提取

在现有的提取特征基因（即基因“标签”）的技术中，大多数采用支持向量机方法（SVM），这是一种非常有效的分类方法。SVM 技术是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力。此方法在解决小样本、非线性及高维模式识别中也表现出许多其特有的优势。本文利用基于灵敏度分析的支持向量机来筛选对于判断是否患有癌症或者肿瘤有重要贡献度的目标基因，即提取基因“标签”。

因为灵敏度函数可以看成是评价某基因对分类决策的影响度的重要性指标，因此可以去除灵敏度比较低的若干基因，得到候选特征基因集合，再通过学习建立新的 SVM 模型，重新计算候选特征基因集合中各基因的灵敏度，去除灵敏度低的若干基因，作为更新的候选特征基因集合，如此反复，直到测试错误增大为止。这样就是利用 SVM 分类器检测各个特征基因集合的分类性能，选择基因数较少，分类性能最佳的基因集合作为与致癌有关的基因“标签”集。

为了讨论的方便，引进如下符号和函数：

给定样本集 $S_T = \{(\vec{x}_i, y_i) \mid \vec{x}_i \in R^d, y_i \in \{-1,+1\}, i = 1, \cdots N\}$

SVM 的判别函数

$$g(x)=sign\left(\sum_{i=1}^{N}a_i y_i k(\vec{x},\vec{x}_i)+b\right) \tag{5.1}$$

分类决策函数

$$O(\vec{x})=\sum_{i=1}^{N}a_i y_i k(\vec{x},\vec{x}_i)+b \tag{5.2}$$

其中，a_i 为 Lagrange 乘子，b 为超平面截距，$k(\vec{x},\vec{x}_i)$ 是核函数。在基因分类中，通常核函数取径向基函数：

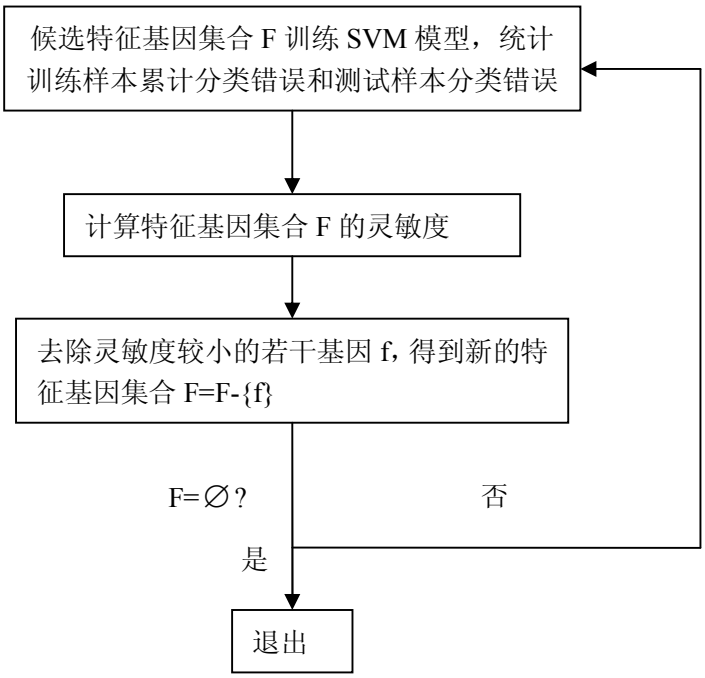
$$k(\vec{x},\vec{x}_i)=e^{\frac{-\|\vec{x}-\vec{x}_i\|^2}{2\sigma^2}} \tag{5.3}$$

分量 x_j 对分类决策函数 $O(\vec{x})$ 的灵敏度函数定义为：

$$S(x_j)=\sum_{X\in S_T}\left|\frac{\partial O(\vec{x})}{\partial x_j}\right|=\frac{1}{\sigma^2}\sum_{\vec{x}\in S_T}\left|\sum_{i=1}^{N}a_i y_i k(\vec{x},\vec{x}_i)(x_{ij}-x_j)\right| \tag{5.4}$$

(5.4)式中 $k(\vec{x},\vec{x}_i)$ 为径向基函数，x_j 为 $\vec{x}$ 的第 j 个分量，x_{ij} 为 $\vec{x}_i$ 的第 j 个分量。

由此我们获得特征基因集提取算法流程如下图(2)所示：



图（2）特征基因集提取算法流程图

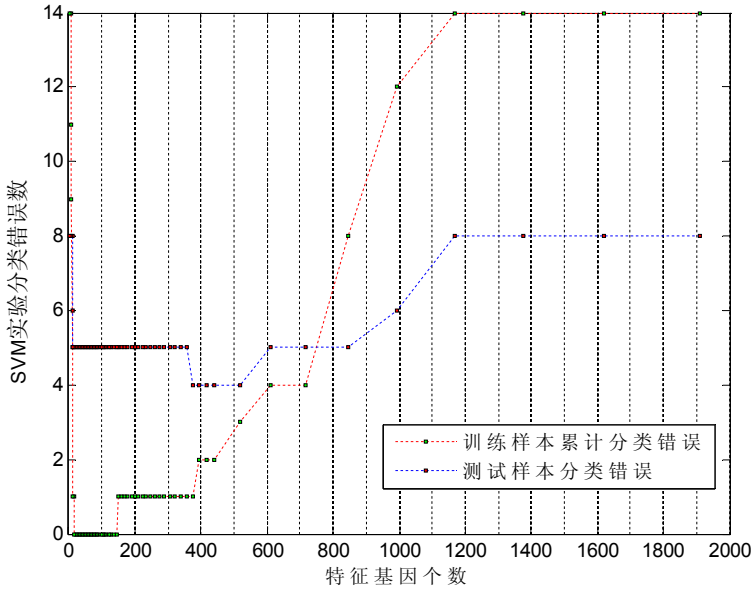
在题目中所给定的 62 个数据样本中，正常人的样本数为 22，癌症患者的样本数为 40，我们选取 40 个样本为训练样本(其中正常人样本数 14，癌症患者样本数为 26)，22 个样本为测试样本(其中正常人样本数为 8，癌症患者的样本数位 14)，具体如下表所示：

表(1)训练和测试样本的选取

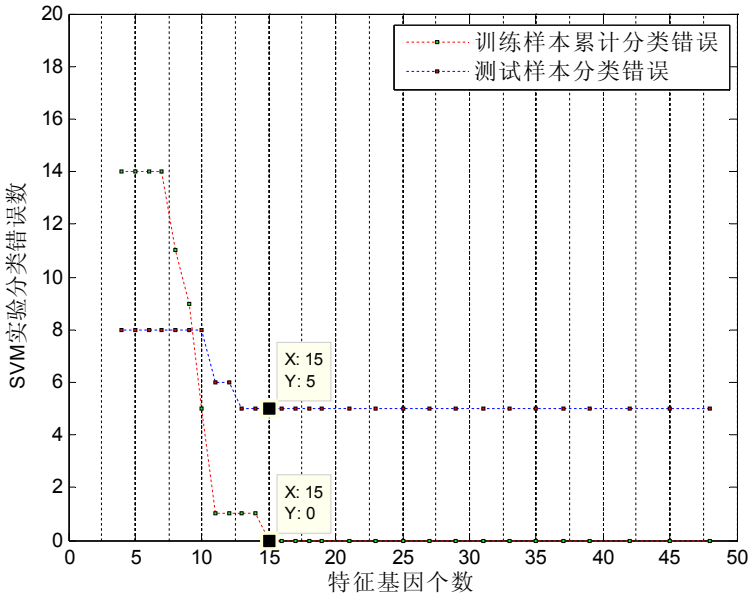
训练样本(总计 40)	测试样本(总计 22)
正常：normal1~normal14	正常：normal15~normal22
癌症病人：cancer1~ cancer26	癌症病人：cancer27~ cancer40

运用前面所述的特征基因提取算法，我们给出在不同的特征基因个数下所产

生的测试样本分类错误数和训练样本累计分类错误数的关系，如下图所示；



图（3）基因与分类能力的关系图



图（4）基因与分类能力的关系图(局部)

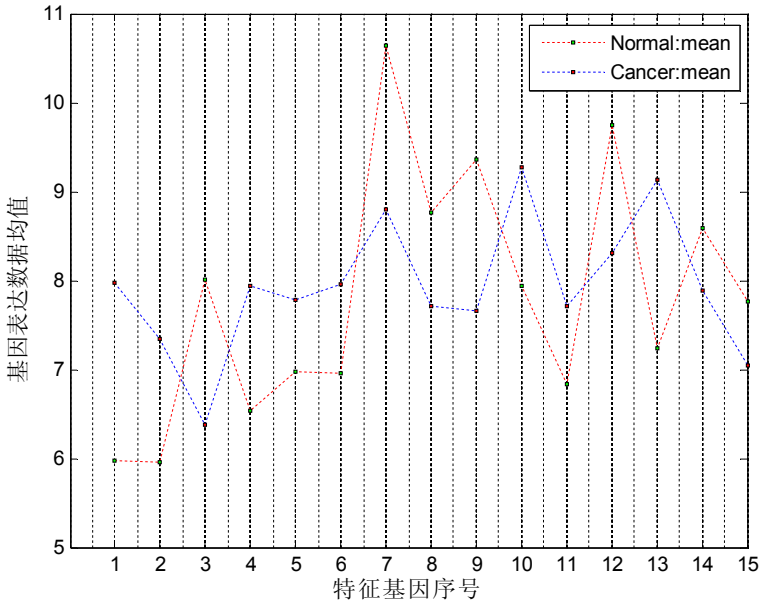
通过上图(3)和(4)可以看出，通过灵敏度对基因进行筛选时，当特征基因数为 15 时，对应的训练样本累计分类错误数和测试样本分类错误数同时达到最小，这就说明 15 是分类性能最佳而特征基因数最少的平衡点。显然，依据这 15 个目标基因就可以比较准确地对测试者是否患有癌症进行判断，也就是基因“标签”。此时筛选出的的 15 个基因“标签”如表(2)所示：

表(2)SVM 提取的基因标签信息

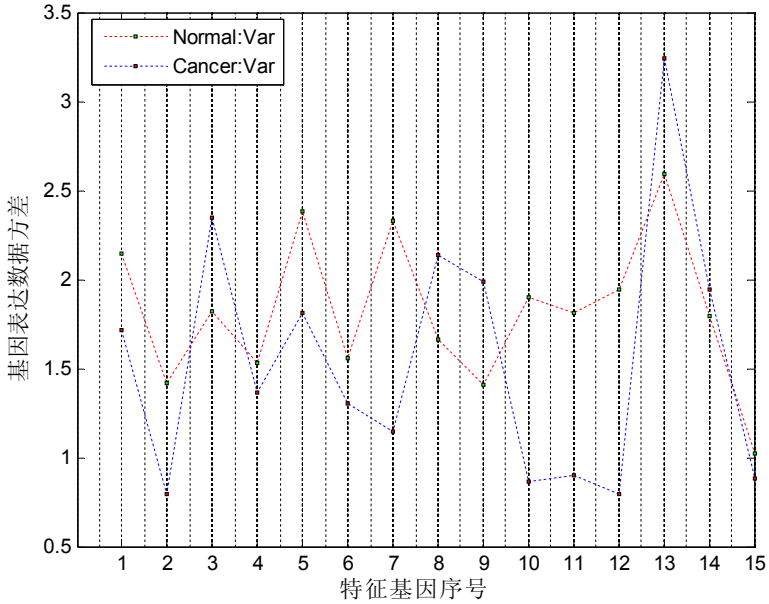
基因描述 / 基因序号	The EST names	GenBank accession ID	Type of region mapped by the EST (3'UTR or gene)	General gene name
基因 1	Hsa.627	M26383	gene	IL8

基因 2	Hsa.601	J05032	gene	DARS
基因 3	Hsa.2291	H06524	3' UTR	GSN
基因 4	Hsa.2645	X54942	gene	CKS2
基因 5	Hsa.108	D13665	gene	POSTN
基因 6	Hsa.229	U04953	gene	IARS
基因 7	Hsa.8147	M63391	gene	
基因 8	Hsa.1387	U14631	gene	HSD11B2
基因 9	Hsa.36952	H43887	3' UTR	CFD
基因 10	Hsa.1047	R84411	3' UTR	SNRPB
基因 11	Hsa.45446	H87135	3' UTR	C7orf47
基因 12	Hsa.692	M76378	gene	
基因 13	Hsa.3016	T47377	3' UTR	S100P
基因 14	Hsa.1280	X16354	gene	CEACAM1
基因 15	Hsa.579	M80815	gene	

对样本中癌症患者和正常人在这 15 个基因上分别计算均值和方差如下图所示：



图（5）癌症患者和正常人在基因标签上的均值



图（6）癌症患者和正常人在基因标签上的方差

从图(5)和图(6)中可以看出，对于提取出的这 15 个基因“标签”的均值和方差在正常人和癌症患者之间有着较大的差距，从而验证了我们采用灵敏度分析的支持向量机进行特征基因的提取准确性。通过查找临床医学研究领域的相关资料发现，在参考文献[11]中 Min Ah Kang 等人的工作已经验证出基因 15 Hsa.579 是与癌症相关的致病基因，而参考文献[9]中吕满义等人在临床学基础上发现基因 8 Hsa.1387 与癌症的关系。由此更加证明此方法获得的基因“标签”是有一定准确度的。

由此，在初步处理后的 1908 个基因中，就得到了少量的基因其表现对是否患病的分类有较高的影响度，而大部分基因的表现对判断是否患有肿瘤或者癌症的贡献度比较小，可以根据筛选出得到的这 15 个基因“标签”就可以对测试者进行正确诊断。

六、基于ICA与SVM的基因标签提取

通过前面所述已经知道，基因图谱中的基因之间存在着很强的关联，将基因簇（基因组）作为区分正常人和癌症患者的分类因素，会更加符合生物医学的理论和事实。本文已经利用独立成分分析(ICA)技术对基因采样数据进行分析，最大程度上降低基因之间强烈的相互影响，将基因信号转化为基因分类信号，从而获得与癌症有关的最有直接关系但数目最少的潜在因素，即基因簇矩阵 $\hat{S}$ 。

在已得到的基因簇的基础上进行分析而得到的与肿瘤或者癌症直接相关的目标基因会更加准确，避免了直接从统计意义上的分类角度出发而可能导致的在突变基因筛选上的错误。

为了在基因簇矩阵 $\hat{S}=(\vec{\hat{s}}_1,\vec{\hat{s}}_2,...\vec{\hat{s}}_n)^T$ 的基础上获得目标基因，本文引入支持向量机方法中的分类模型：

$$\sum_i^n C_i \hat{S}_{i,t} = \begin{cases} 1 \\ -1 \end{cases}$$

其中， $t=1,2,\ldots,sn$ ， sn 为样本数。由于 $\hat{S}=WX$ ，显然

$$\sum_i^n C_i \hat{S}_{i,t} = \sum_i^n C_i (WY)_{i,t} = \sum_i^n (\bar{C}W)_i Y_{i,t}$$

那么分类模型就是

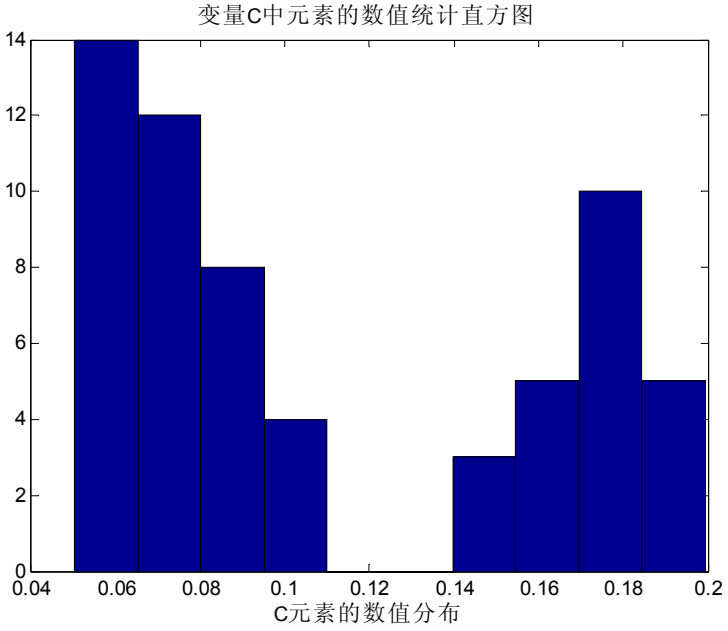
$$\sum_i^n C_i \hat{S}_{i,t} = \sum_i^n (\bar{C}W)_i Y_{i,t} = \begin{cases} 1 & t \in normal \\ -1 & t \in cancer \end{cases} \quad (6.1)$$

这样就构建了一个关于向量$C_i(i=1,2,\ldots n)$的方程组。向量$C_i(i=1,2,\ldots n)$的取值就显示了对应基因簇（基因组）$\vec{\hat{S}}_i(i=1,2,\ldots,n)$对分类的贡献程度，而$(CW)_i$的取值则显示了由基因簇为基础的基因对分类的影响程度，影响程度越高，对应基因就越可能是与癌症致病有关的基因。

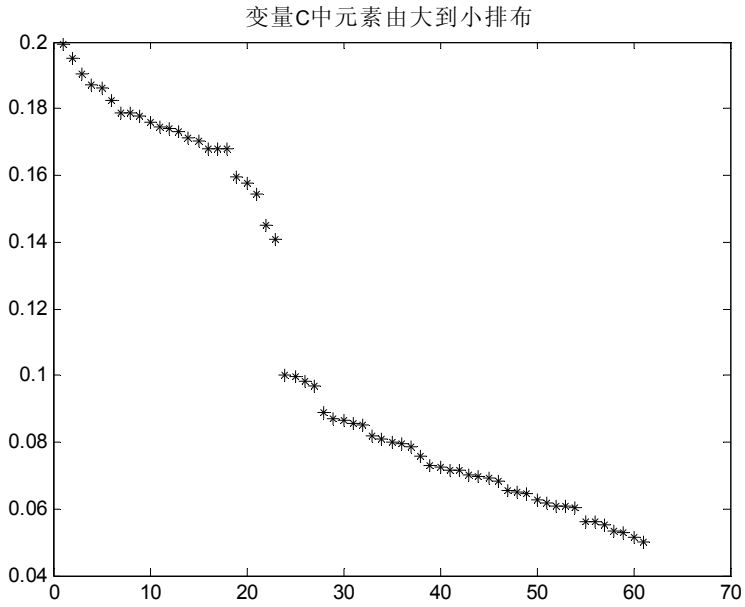
众所周知，与癌症相关联的基因潜在因素或蛋白质信号数目很少。能够根据较少量的蛋白质信号就可以以上进行分类，所以向量C就要提高稀疏度，也就是说较多的元素趋近于零。因此，目标函数可以写为：

$$f(\bar{C}) = \min \left\| \bar{C} \right\|^p \qquad 0 < p < 1 \quad (6.2)$$

由此就形成了一个非线性等式约束问题。其实，对于解非光滑优化问题是非常困难的，这一问题也是目前研究的一个热点，尚未得到很好的解决，这也是需要今后我们继续研究的问题。本文中选取$p=0.2$，可计算得到向量$\vec{C}$的解，此解只是局部最优解，其数值分布如下图(7)、图(8)所示，从图中可以看到得到的局部最优解还是有比较强的稀疏性。不过，如果能得到问题的全局最优解一定更加稀疏。



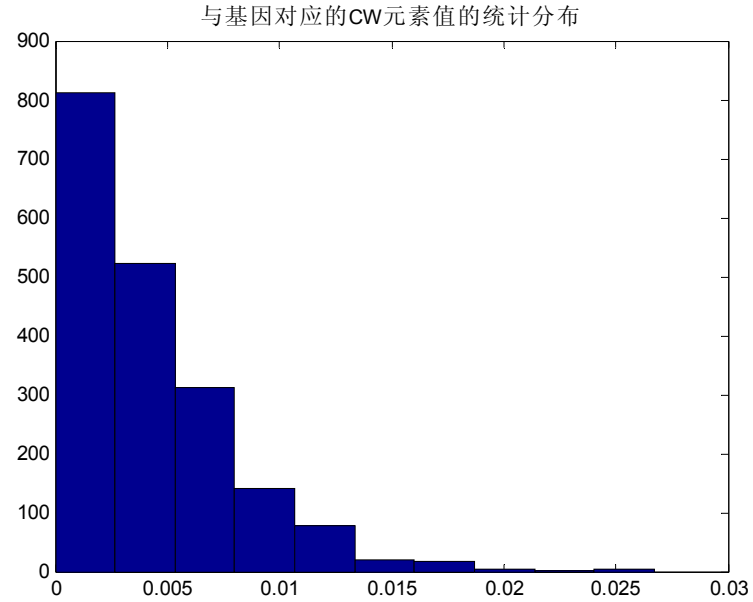
图（7）向量 C 中元素的数值直方图



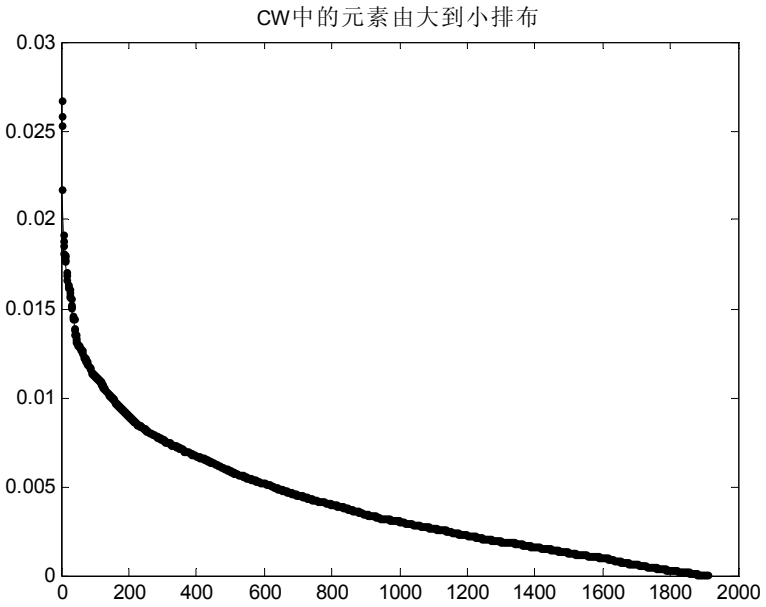
图（8）向量 C 中元素由大到小的排布图

通过统计，有 23 个元素集中在[0.14,0.2]的范围内，38 个元素集中在[0.05,0.1]的范围内，而且大部分分布在小于 0.09 的值域内。从图(8)中，可以明显的看到，在[0.1,0.14]范围内出现了较大的断点，这就说明此向量 $\bar{C}$ 是具有一定稀疏性的。[0.14,0.20]的范围内的元素代表比较大的权重，也就是说对于分类具有较大的贡献。而[0.05,0.1]的范围内的元素代表影响度很低的，因此可以将此部分元素值直接设置为 0，降低这些低贡献度的影响因子在后续的致病有关基因的选择中造成干扰。

将此计算出的向量 $\bar{C}$ 与独立成分分析中得到的分离矩阵 W 相乘，就可以得到与基因对应的分类贡献度 $\bar{C}W$ ，其具体分布如下图：



图（9）向量 CW 中元素的数值直方图



图（10）向量 CW 中元素由大到小的排布图

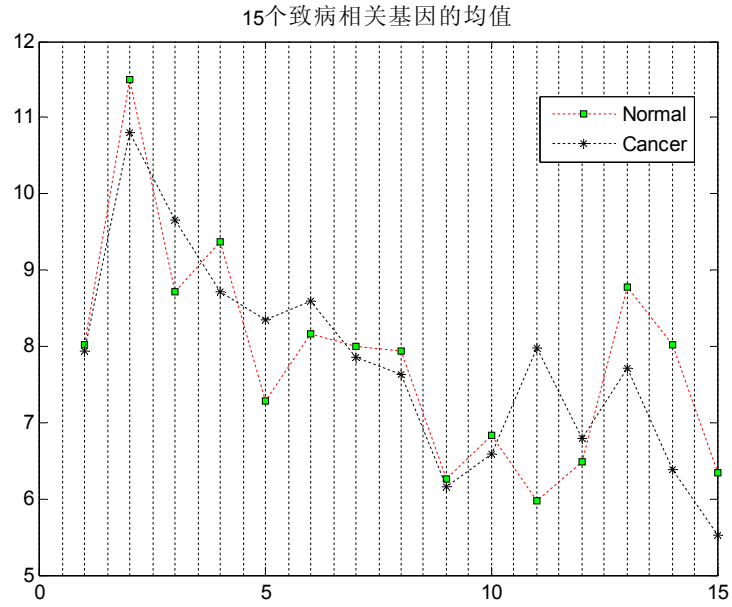
据统计，在与基因对应的分类贡献度向量 $\bar{C}W$ 中，只有 32 个元素权重在 0.015 以上，而权重在 0.01 以下的有 1759 个元素，其中有 1275 个元素权重在 0.005 以下。也就是说，在初步处理后的 1908 个基因中，只有少数基因的表现对是否患病的分类有较高的影响度，也就是说这些基因可能是致病基因或者是亚致病基因，而大部分基因的表现与判断是否患有肿瘤或者癌症没有显著的直接关系。

由此选出影响度最高的前 15 个基因作为与癌症紧密相关的基因“标签”，如下表：

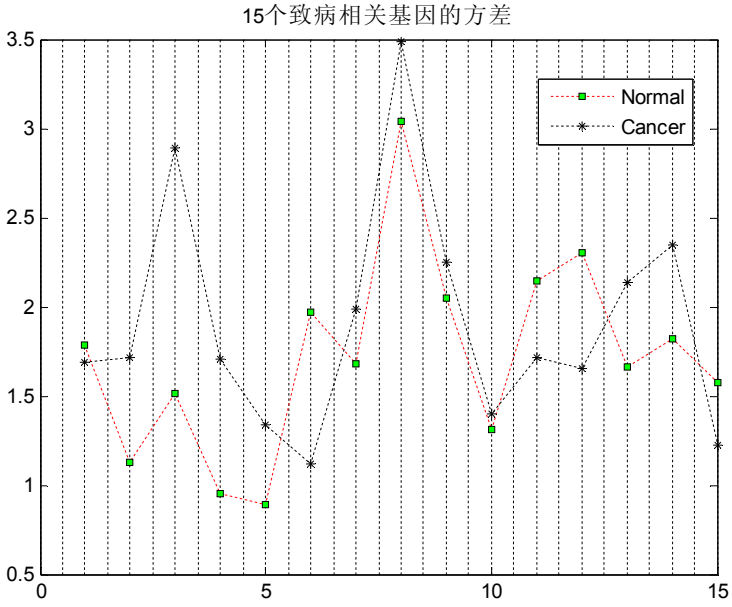
表（3）15 个基因“标签”信息

基因描述 / 基因序号	The EST names	GenBank accession ID	Type of region mapped by the EST (3'UTR or gene)	General gene name	在 CW 中对应的权重值
基因 1	Hsa.2918	X67325	gene	IFI27	0.0267
基因 2	Hsa.1737	T72175	3' UTR		0.0258
基因 3	Hsa.891	M19045	gene	LYZ	0.0253
基因 4	Hsa.2809	R70030	3' UTR		0.0217
基因 5	Hsa.9972	T51261	3' UTR		0.0192
基因 6	Hsa.2135	U21128	gene	LUM	0.0188
基因 7	Hsa.3353	X02492	gene	IFI6	0.0188
基因 8	Hsa.2706	X02761	gene	FN1	0.0185
基因 9	Hsa.27808	R51502	3' UTR	SF3B4	0.0181
基因 10	Hsa.40211	R99935	3' UTR	PAPSS2	0.0181
基因 11	Hsa.627	M26383	gene	IL8	0.0179
基因 12	Hsa.8121	H11125	3' UTR	CCND2	0.0178
基因 13	Hsa.1387	U14631	gene	HSD11B2	0.0177
基因 14	Hsa.2291	H06524	3' UTR	GSN	0.0171
基因 15	Hsa.72	D29808	gene	TSPAN7	0.0171

癌症患者和正常人在这 15 个基因“标签”上的均值和方差如图：



图（11）癌症患者和正常人在基因“标签”上的均值



图（12）癌症患者和正常人在基因“标签”上的方差

从图(11)和图(12)中可以发现，癌症患者和正常人在这 15 个基因上的均值和方差并不是都有明显的差别，例如表(3)中的基因 1、基因 7、基因 9、基因 10。

通过查找临床医学研究领域的相关资料发现，在参考文献[7]中邹琼，朱道奇等人的研究工作已经发现基因 1 Hsa.2918 是与癌症相关的致病基因，参考文献[8]中金钢，胡先贵等人证实基因 8 Hsa.2706 与癌症有着密切的关系，参考文献[9]中吕满义等人在临床学基础上发现基因 8 Hsa.1387 与癌症之间也存在着重要关系，而参考文献[10]中 Xinan Yang 等人研究得出基因 15Hsa.72 也是关键的癌症致病基因。由此更加证明此方法获得的基因“标签”是有一定的精确度的，而且从某种程度上说明基于基因簇的 ICA 与 SVM 相结合的基因“标签”提取方法更加出色。

这些基因是从致病有关的基因簇（也可以说是蛋白质信号）的角度出发进行的筛选，而并非基因本身的表现，这一思路比较新颖。从蛋白质信号的角度出发，更加贴近生物医学的事实，更深入的探索了基因之间的联系与是否导致癌症之间的复杂关系，由此判断出来的致病有关基因“标签”可能会更加有生物医学上的意义。

与单从基因表现上进行筛选的 SVM 方法得到的 15 个基因“标签”进行对比，有 3 个基因是相同的，分别是：

表(4)两种方法获得的相同基因“标签”信息

基因描述 / 基因序号	The EST names	GenBank accession ID	Type of region mapped by the EST (3'UTR or gene)	General gene name
基因 1	Hsa.627	M26383	Gene	IL8
基因 2	Hsa.1387	U14631	gene	HSD11B2
基因 3	Hsa.2291	H06524	3' UTR	GSN

这说明无论是从蛋白质信号或者基因簇的角度出发还是从基因本身的表现角度出发，这 3 个基因都是和患有癌症或者肿瘤有较大关系的，是判断是否患病的最显著的目标基因。

七、基于噪声模型的基因标签提取

众所周知，通常基因图谱数据含有三种噪音：1.系统噪声，基因探测仪器的芯片问题导致数据的不准确，例如数据预处理中提到的重复测量和探针的错误探测；2.无关基因导致的噪声，噪声是一种相对的概念，对于识别问题，与识别无关的基因都可以归纳为噪声；3.随机测量噪声，多种因素导致测量数据与真实值有一定的偏差。此外，并不排除数据中存在错误诊断和基因在肿瘤或者癌肿的不同时期的不同表现，这些也将对判断与致病有关的目标基因造成影响。事实上，肿瘤是一个发展过程，将不同阶段的数据放在一起考虑可能会产生误差。

在本文数据预处理中，已经删除了部分系统噪声，但是仍然存在其他噪声的影响。特别是随机误差与无关基因。通常，我们总假设随机误差的值很小，但正如文献[3]的工作所指出的那样，有时有些基因的表达水平很低，很容易淹没在噪声之中。这时，如果直接采用 PCA 等方法，会使得一些基因“标签”无法判断。所以，在我们的分析中需要建立含有噪声的模型。

在前面的讨论中，我们已经利用 ICA 得到了好的分类因素，不过没有考虑噪声因素。下面，我们在独立成分分析时考虑到噪声因素，由此构建含有噪声 N 的独立成分分析模型：

$$X = AS + N \tag{7.1}$$

式中 N 是噪声，S 由两部分组成：分类因素和无关因素。我们知道噪声是很难估计的，其实并没有估计价值。所以，我们将上式改写成：

$$X = A(S + N) \tag{7.2}$$

即通过独立成分分析得到的独立成分是真实基因组合和噪声的某种组合，为了简单其间，我们依然用 N 来表示噪声。

根据前面的工作，通过 ICA，我们可以得到估计：$S + N$ 。由于并不是真正

的分类因素，所以由此进行分类会出现误差，因此在构造分类器时，我们增加了松弛因子，即相应的分类模型也需要增加模糊性：

$$\sum_{i}^{n} C_i (S+N)_{i,t} = \begin{cases} 1+\xi & t \in normal \\ -1+\xi & t \in cancer \end{cases} \tag{7.3}$$

其中ξ是松弛因子。

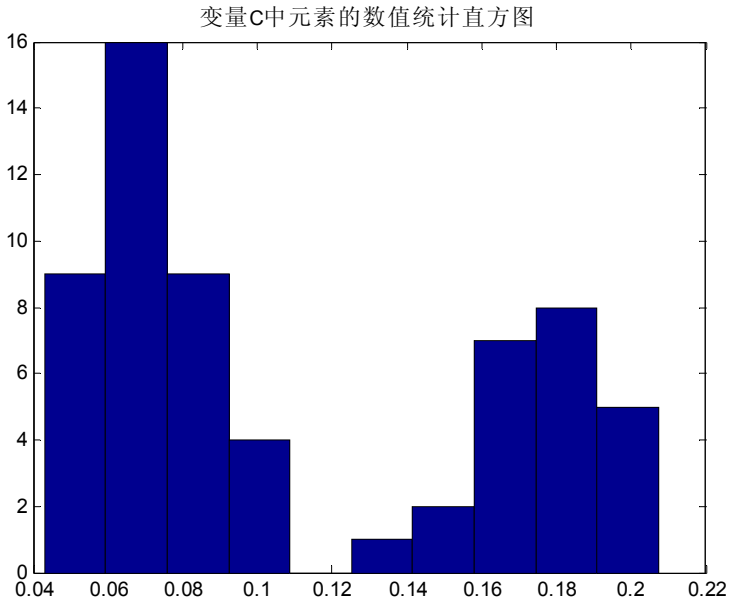
本文中，$\xi = \pm 0.5$，这样就在一定程度上增加了分类的模糊性，也就是考虑到噪声对分类的影响。

另一方面，由于通过 ICA 得到的潜在因素里，含有无关因素。为了能够去除其对分类的影响，数学上只要要求这些因素前面的稀疏为零。同时，我们知道通过潜在因素的确定，用于分类的因素会很少，即上述公式中系数的零元素达到最大。为此，我们希望C要提高稀疏度。由此就形成了一个非线性等式约束问题：

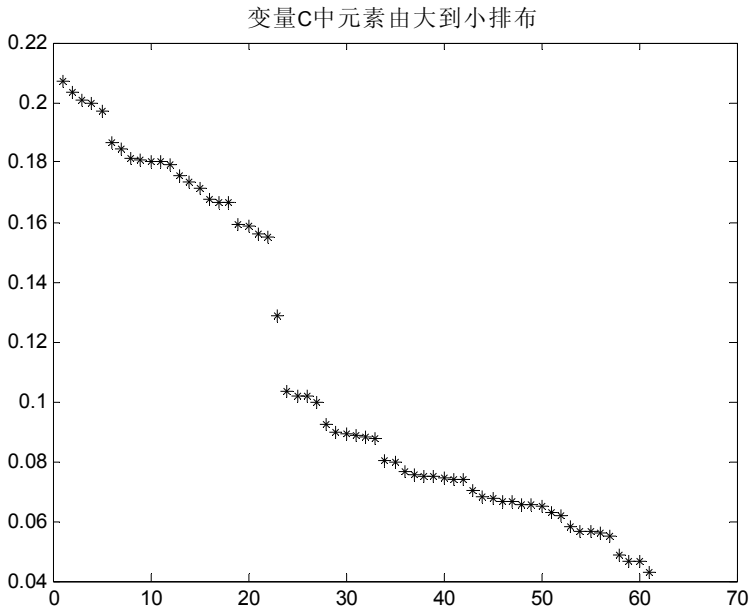
$$f(\vec{C}) = \min \left\| \vec{C} \right\|^{p} \qquad 0 < p < 1$$

$$\text{s.t.} \qquad \sum_{i}^{n} C_i (S+N)_{i,t} = \begin{cases} 1+\xi & t \in normal \\ -1+\xi & t \in cancer \end{cases} \tag{7.4}$$

本文中取值$p = 0.2$，$\xi = \pm 0.5$，可计算得到向量$\vec{C}$的局部最优解解，其数值分布如下图：



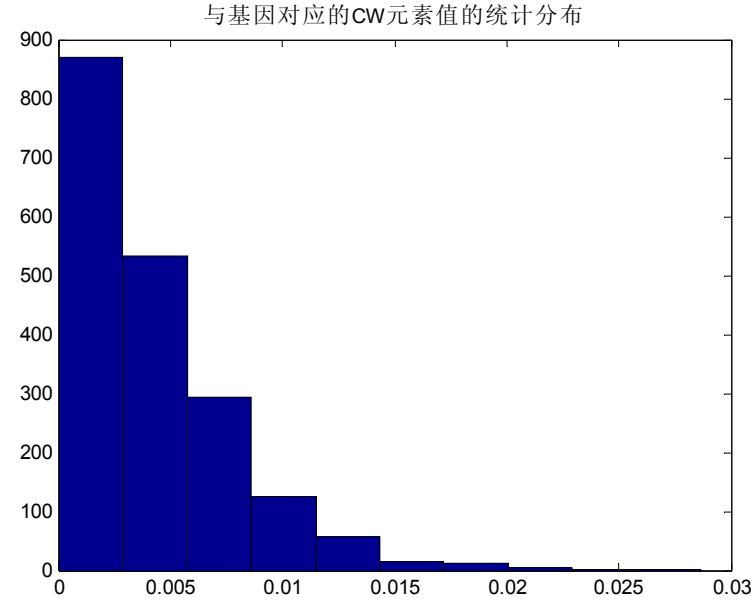
图（13）向量 C 中元素的数值直方图



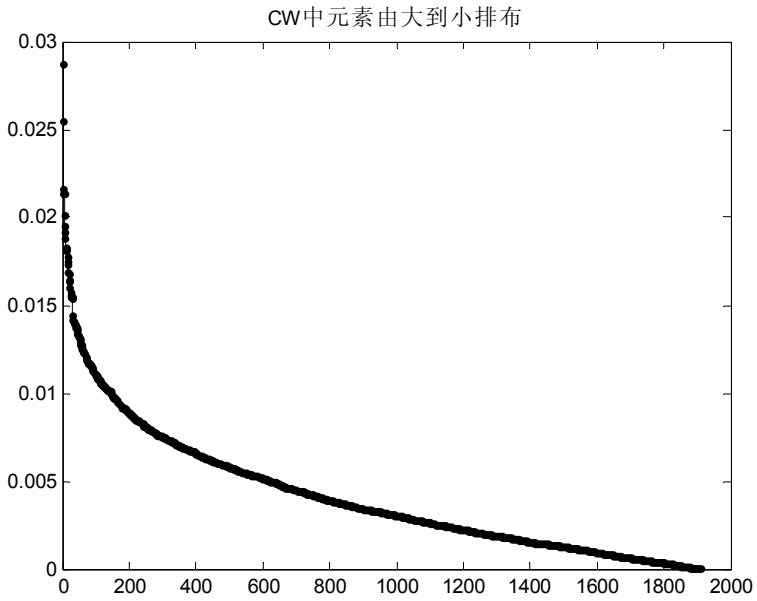
图（14）向量 C 中元素由大到小的分布图

通过统计，有 22 个元素集中在[0.15,0.21]的范围内，只有一个点在 0.28 左右，38 个元素集中在[0.04,0.11]的范围内，而且大部分的点分布在小于较小 0.09 的值域内。从图中，可以明显的看到，在[0.11,0.15]范围内出现了较大的断点，这就说明此向量$\bar{C}$是具有一定稀疏性的。在[0.15,0.20]范围内的元素代表比较大的权重，也就是说对于分类具有较大的贡献。而其它的元素代表影响度很低，仍然将此部分元素值直接设置为 0，降低这些低贡献度的影响因子在后续的致病有关基因的选择中造成干扰。

将此计算出的向量$\bar{C}$与独立成分分析中得到的分离矩阵W相乘，就可以得到与基因对应的分类贡献度$\bar{C}W$，其具体分布如下图(15)、图(16)：



图（15）与基因对应的 CW 元素值的直方图



图（16）与基因对应的 CW 元素由大到小分布图

据统计，在与基因对应的分类贡献度向量 $\bar{C}W$ 中，只有 29 个元素权重在 0.015 以上，而权重在 0.01 以下的有 1762 个元素，其中有 1282 个元素权重在 0.005 以下。仍然说明，在初步处理后的 1908 个基因中，只有少数基因的表现对是否患病的分类有较高的影响度，而大部分基因的表现与判断是否患有肿瘤或者癌症没有直接的关系。

对比之前没有噪音的结果：

表（5）有无噪声模型结果的对比表

	整体取值范围	大于 0.015	小于 0.01	小于 0.005
无噪声模型	[0,0.027]	32	1759	1275
有噪声模型	[0,0.029]	29	1762	1282

显然，考虑到噪声影响的非线性等式约束模型得到的与基因对应的分类贡献度向量 $\bar{C}W$ 更加稀疏，对判断是否患病有重要贡献的基因的权重更加突出，由于噪声影响而导致权值取值有误的基因被排除，由此降低了噪声对寻找基因“标签”的影响。由此选出影响度最高的前 15 个基因，即基因“标签”，如下表：

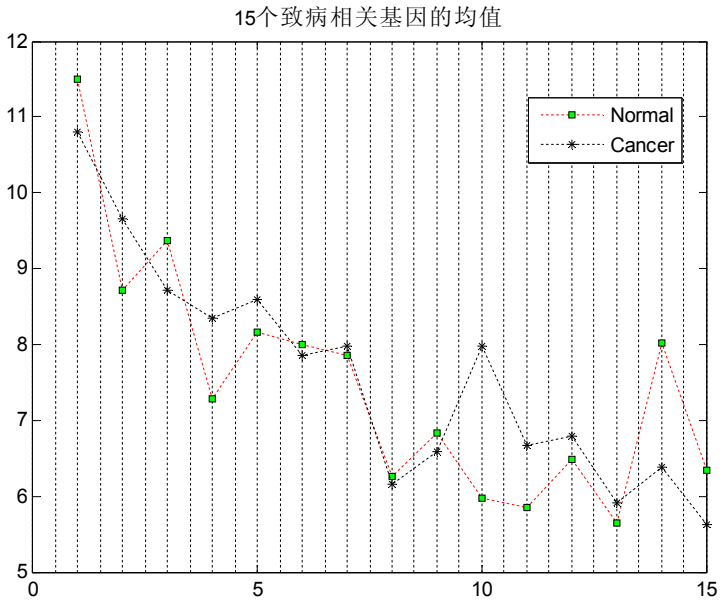
表（6）15 个基因标签信息

基因描述 / 基因序号	The EST names	GenBank accession ID	Type of region mapped by the EST (3'UTR or gene)	General gene name	在 CW 中对应的权重值
基因 1	Hsa.1737	T72175	3' UTR		0.0287
基因 2	Hsa.891	M19045	Gene	LYZ	0.0255
基因 3	Hsa.2809	R70030	3' UTR		0.0216
基因 4	Hsa.9972	T51261	3' UTR		0.0214
基因 5	Hsa.2135	U21128	Gene	LUM	0.0214
基因 6	Hsa.3353	X02492	Gene	IFI6	0.02101
基因 7	Hsa.2922	X69910	Gene	CKAP4	0.0195
基因 8	Hsa.27808	R51502	3' UTR	SF3B4	0.0192

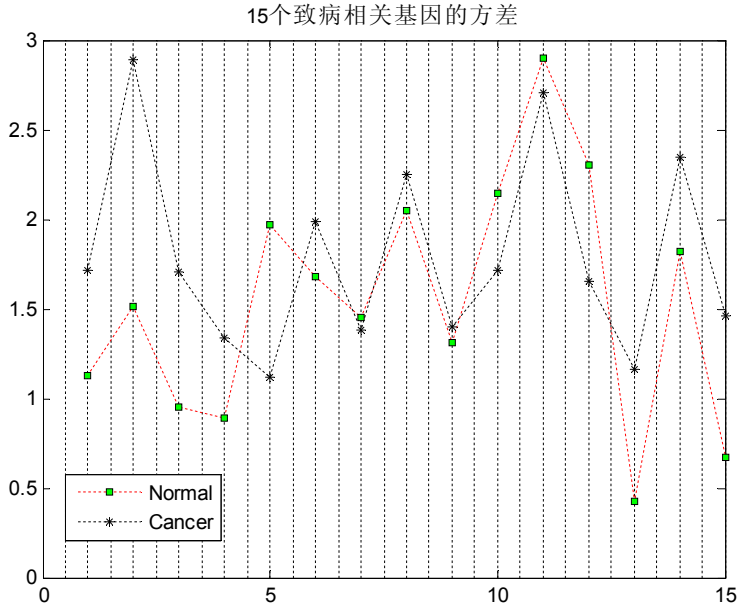
基因 9	Hsa.40211	R99935	3' UTR	PAPSS2	0.0188
基因 10	Hsa.627	M26383	Gene	IL8	0.0183
基因 11	Hsa.38171	R89823	3' UTR		0.0182
基因 12	Hsa.8121	H11125	3' UTR	CCND2	0.0182
基因 13	Hsa.2012	M81651	Gene		0.0181
基因 14	Hsa.2291	H06524	3' UTR	GSN	0.0178
基因 15	Hsa.43331	H64807	3' UTR		0.0175

其中有 4 个基因和无噪声的模型得到的基因“标签”是不同的，分别是 Hsa.2922、Hsa.38171、Hsa.2012、Hsa.43331，11 个基因“标签”是相同的。

癌症患者和正常人在这 15 个基因“标签”上的均值和方差如图：



图（17）癌症患者和正常人在基因标签上的均值

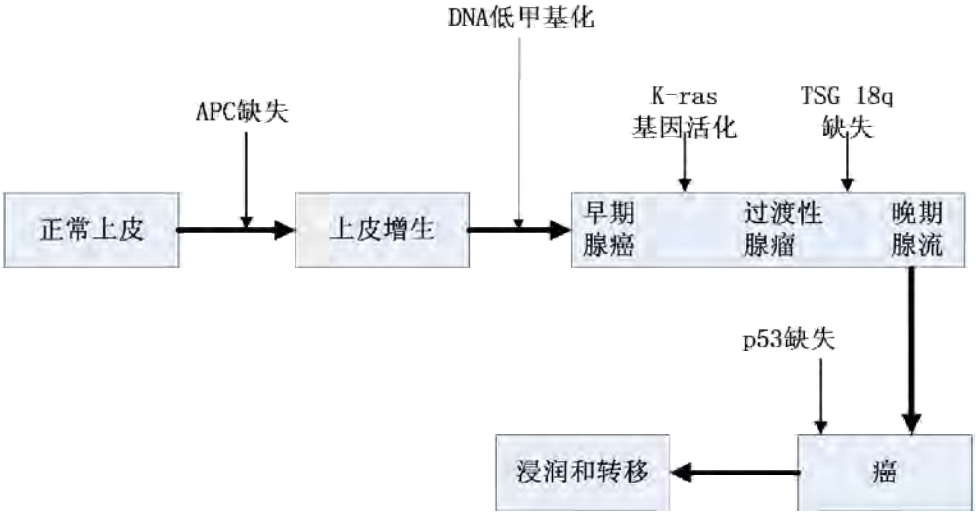


图（18）癌症患者和正常人在基因标签上的方差

从图(17)、图(18)上可以发现，癌症患者和正常人在这 15 个基因“标签”上的均值和方差仍然并不是都有明显的差别，例如基因 7 和基因 9。但是，方差上的差距整体上比不考虑噪声而选择的基因的方差要明显一些。

八、结合临床结论的数学模型

众所周知，在肿瘤研究领域的临床学上已证实若干个基因的变异是导致癌症的重要因素，在提取的特征基因的过程中，需要融入这些明确的癌症相关基因信息。临床学中有以下发现：大约有 90%的结肠癌在早期有 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。癌生物学研究证实结肠癌的发展过程如下：



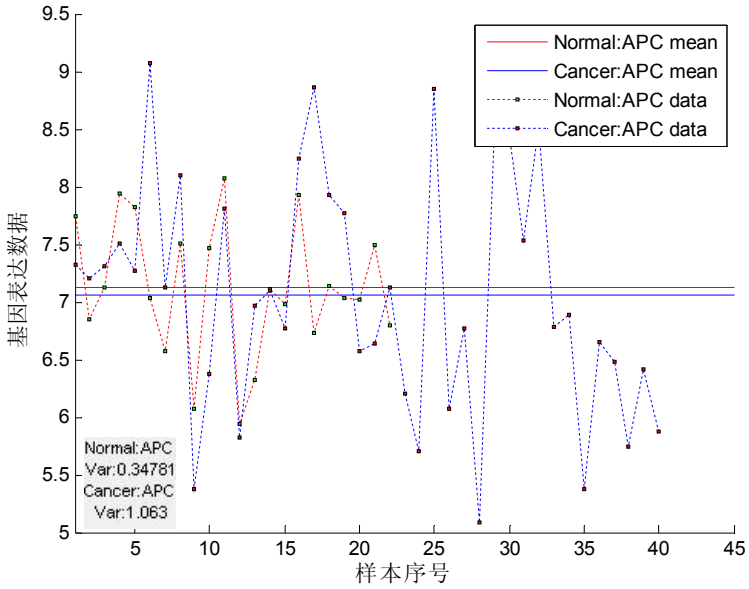
图（19）癌症发展过程图

一般地，我们认为正常样本与患病样本的 APC 基因与 Ras 基因的表达水平是有明显不同的，样本均值与方差可以体现基因表达水平的差距。从原始数据中我们可以找出一个 APC 基因和 7 个 ras 基因。进过分析，根据具有较大的基因表达差距来筛选出了 APC、ras 基因各一个，如表（7）所示。

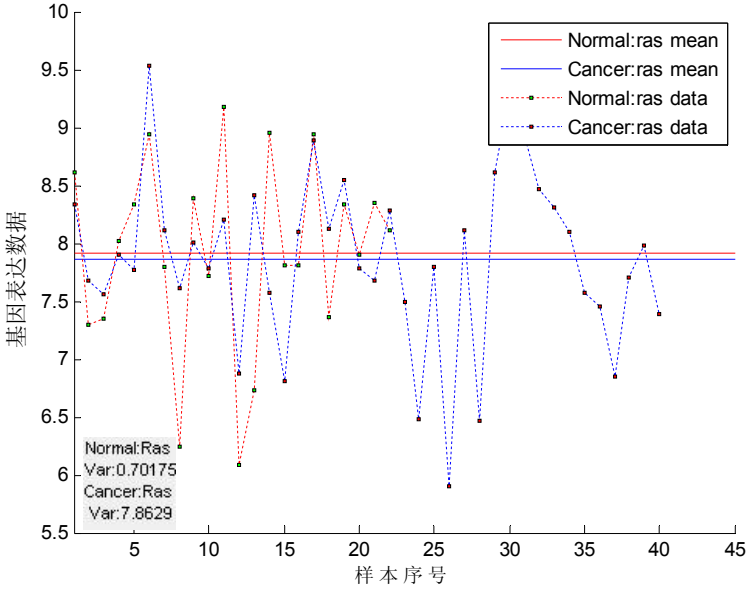
表（7）APC 基因、K-ras 基因信息

基因描述 / 基因序号	the EST names	GenBank accession ID	Type of region apped by the EST (3'UTR or gene)	A brief description of the corresponding gene	General gene name
APC	Hsa.2238	L35545	gene	"Homo sapiens endothelial cell protein C/APC receptor (EPCR) mRNA, complete cds	PROCR
Ras3	Hsa.3909	R53941	3' UTR	RAS-RELATED C3 BOTULINUM TOXIN SUBSTRATE 1 (Homo sapiens)	

原始数据中的 62 个样本（22 个正常，40 个非正常）在以上 APC 基因与 Ras 基因上的统计信息如图(20)、图(21)。



图（20）APC 基因的数据统计



图（21）Ras 基因的数据统计

事实上，由于只有大约 90%的结肠癌病人的 APC 基因失活，说明还有约 10%的结肠癌病人的 APC 基因没有突变，这就给基因分类带来的混乱。所以，我们可以利用这个条件对病人样本进行筛选，使得剩余的病人样本 APC 基因几乎都出现突变。同样对于 ras 基因的变异也进行样本的过滤。

这样在已知临床信息的基础上进行基因筛选，可以对应于数学上的条件概率问题。如果用条件概率语言阐述上述事件，设 A 表示结肠癌在早期有 5 号染色体长臂 APC 基因的失活， B 表示 K-ras 基因突变， C 表示筛选得到的标签基因，则上述事件为条件概率 $P(C|A,B)$ 。

对于 40 个癌症患病样本，我们只希望获取 APC 基因失活的那部分样本，这样得到的分类准确率更加精确。若其 APC 基因表达水平与正常样本的 APC 基因表达水平的均值近似相等，则认为其 APC 基因是正常表达的，所以有必要去除

这些样本。在此基础上，我们还需去除 ras 基因表达水平与正常 ras 表达水平均值差别不大的样本，即去除 ras 基因正常表达的样本，最终得到一个样本子集。

对于 APC 基因，不妨设定一个较小的阈值 $\lambda = 0.1$，定义

$$\hat{\mathrm{X}}_{APC}(i) = \begin{cases} \mathrm{X}_{APC}(i) & 若 \left|\mathrm{X}_{APC}(i) - \Psi_{APC}\right| < \lambda \\ \varnothing & 其它 \end{cases} \quad (i = 1,2,\cdots 40) \qquad (8.1)$$

其中，$\mathrm{X}_{APC}(i)$ $(i = 1,2,\cdots 40)$ 为癌症患病样本的 APC 基因表达水平，Ψ_{APC} 为正常样本的 APC 基因表达水平的均值，$\hat{\mathrm{X}}_{APC}$ 为筛选得到的癌症患病样本集。

根据(8.1)式，本文中共去除了 4 个样本。与 APC 基因类似，在已筛选的样本基础上，设定新的阈值 $\lambda = 0.2$，按 ras 基因重新筛选，最终得到的患病样本数为 26。这样，总样本数由 62 变为 48（26 个癌症患病样本，22 个正常样本）。

我们利用 ICA 得到的与癌症有关的基因“标签”是不包括 APC、ras 基因的。现在在 ICA 提取出来的 15 个与癌症有关的基因“标签”的基础上，加上 APC、ras 基因，即目标基因集合包含 17 个特征基因。对包含 17 个与癌症有关的目标基因的 48 个新样本，选择不同的训练样本数(从 5 到 15)与测试样本数(从 5 到 15)，利用 SVM 进行样本分类，并比较包含 APC、ras 基因和不包含 APC、ras 基因的分类性能，

经过统计我们发现在所做的 121 组测试中，如果在加上 APC、ras 基因的情况下，有 38 组的测试样本识别率优于不加入 APC、ras 基因的测试样本识别率，识别率增大值分别在[2.5,14.9]范围内，平均增大了 4.9；有 58 组的测试样本识别率与不加入 APC、ras 基因的测试样本识别率相同；只有 25 组的测试样本识别率比不加入 APC、ras 基因的测试样本识别率低，识别率减小值分别在[2.7,9.4]范围内，平均减小了 4.6。

显然，在考虑到 APC、ras 基因的情况下，针对不同的训练样本数，大部分情况下识别率是高于或者等于无 APC、ras 基因的模型，而且优化数目大于劣化数目，优化程度也大于劣化程度。从而进一步论证了，APC、ras 基因与直肠癌的关系是十分密切的，它们确实是有助于确定诊断直肠癌信息的基因标签。这样，将临床结论与数学方法得到的标签基因相结合的方式的确增加了判断是否患病的正确率，结合了临床医学理论和数学模型分类的不同优点，更有实际应用价值。

九、模型问题及改进

模型中存在的问题以及需要改进的地方：

（1）在 ICA 和 SVM 相结合的模型中用到了解非光滑优化问题，对于这种类型的优化问题，一直比较棘手，暂时还有得到较好的解决，需要今后不断地学习研究，找到更好的算法，建立更好的数学模型进行求解。

（2）生物信息学是当今生命科学和自然科学的重大前沿领域之一，同时也将是 21 世纪自然科学的核心领域之一，其研究重点主要体现在基因组学和蛋白组学两方面。基因表达数据中，不仅关注基因簇，而且需要注重结果的精度与效率，这是基于基因组学方面的考虑。另一方面则需要注意到可能存在具有相似结构和功能的基因簇，因为这更能揭示生物现象的本质，这是基于蛋白质学方面的考虑。由于对蛋白质的复杂结构和形成并不是很了解，还有很多奥秘有待探索研究，这是今后需要进一步研究的方向。

（3）传统的基于统计学习的数据挖掘技术由于相应的样本或是特征较少，缺乏

真正意义上的统计重要性，而需要利用多元信息融合的方法来研究基因表达数据。融合各领域其他重要信息，对基因表达信息进行补充，可以在生物功能发面对获得的结果进行更为合理的解释，增强所得结果在生物医学中的实际作用。这也将是今后努力的方向。

参考文献

[1] Aapo Hyvarinen, Juha Karhunen, Erkki Oja. 独立成分分析[M]. 北京, 电子工业出版社,2007.

[2] Vladimir N. Vapnik, 计学习理论[M] , 北京, 电子工业出版社, 2009.

[3] 李瑶, 因芯片数据分析与处理[M]. 北京, 化学工业出版社, 2006.

[4] 谢纳, 生物芯片分析[M], 北京, 科学出版社, 2004.

[5] Gary Hardiman, 生物芯片技术与应用详解[M]. 北京, 化学工业出版社, 2006.

[6] 刘全金, 李颖新, 阮晓刚. 基于 SVM 的灵敏度分析方法选取肿瘤特征基因[J]. 北京工业大学学报, Vol.33 No.9 Sep.954-957 , 2007.

[7] 邹琼, 朱道奇, 章宗籍, 申丽娟. 肝癌及癌前病变中 p27 蛋白和 mRNA 的表达[J]. 中国肿瘤,第一期.

[8] 金钢, 胡先贵, 应康, 李瑶, 唐榕, 唐岩, 景在平, 谢毅, 毛裕民. 基因表达谱芯片在胰腺癌相关基因筛选中的应用研究[J]. 上海, 第二军医大学学报, Vol.21, No.9, 819-822, 2000.

[9] 吕满义, 纪新强, 高占军, EACAM1 和 VEGF 在宫颈鳞癌组织中的表达及临床意义[J]. 肿瘤基础与临床, 2008 年 21 卷 1 期.

[10] Xinan Yang and Xiao Sun.. novel serum biomarker for pancreatic cancer[J]. Meta-analysis of Cancer Gene-Profiling Data, 2007,04.

[11] Min Ah Kang, Jong-Tae Kim, Joo Heon Kim. Upregulation of the cycline kinase subunit CKS2 increases cell proliferation rate in gastric cancer[J]. Journal of Cancer Research and Clinical Oncology, Volume 135, Number 6, 761-769.

[12] T. R. Golub. Molecular ClassiÞcation of Cancer: Class Discovery and Class Prediction by Gen Expression Monitoring[J]. science, VOL 286, 581-587.

[13] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J., road patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proc Natl Aead Sci Usa, 1999, 96:6745-6750.

[14] V.N.Vapnik. Statistical Learning Theroy[J].New York: Wiley Interscience,1998.

[15] Furey,T.S.,et al.,Support vector machine classification and validation of Cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000,16(10):906-14.

[16] Golub T R,Slonim D K,Tamayo P et al. Molecular classifica-tion of cancer:class discovery and class prediction by gene ex-pression monitoring [J]. Science, 1999, 286: 531-537.

[17] 杨福生, 洪波. 独立分量分析的原理与应用[M]. 北京：清华大学出版社, 2006.

YANG Fu-sheng, HONG Bo. Theory and application of Independent Component Analysis [M]. Beijing：Tsinghua University Press，2006.

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

[19] Lee T-W，Girolami M，Sejnowski T J. Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources [J]. Neural computation (S0899-7667)，1999，11(2)：417-441.

[20] Lu Y, Han J W. Cancer classification using gene expression data[J]. Inform Syst, 2003, 28(4): 243~268