全国第七届研究生数学建模竞赛

题 目　　基于 SVM 和 LDA-GA 的基因图谱信息提取方法的研究

摘　　要：

本文针对提取基因图谱信息的问题，运用浮动顺序搜索算法、RBF 支持向量机和遗传线性判别算法（LDA-GA）等方法，在不处理噪声、降噪以及融入其他有价值的信息三种条件下分别建立能够有效提取样本基因图谱信息的模型，并利用样本数据针对每种条件下得到的基因“标签”的分类能力进行测试和分析。

针对问题 1，首先以 Bhattacharyya 距离为评价函数，对样本中 2000 个基因进行无关基因的剔除，得到 388 个信息基因；然后，在信息基因集合中，根据浮动顺序搜索算法搜索得到 35 个候选分类特征子集，为问题 2 中基因标签的筛选提供必要条件。

针对问题 2，根据样本数据，利用候选分类特征子集对 RBF 支持向量机进行训练，采用“留一法”和“独立测试实验”对所建支持向量机进行测试。通过对测试结果的分析与评价，筛选出具有最佳分类效果的特征子集，作为基因“标签”。通过实验得到的基因“标签”为 7 维向量。

针对问题 3，分析 NT_I 及 NT_Ⅱ两类噪声，建立噪声模型并对样本数据进行降噪处理。运用处理后的样本数据，确定新的基因“标签”。实验结果表明，新的基因“标签”具有更高的分类精度。

针对问题 4，根据有助于诊断肿瘤的相关信息，利用 LDA-GA 方法对有价值的生理基因进行筛选得到最优生理基因向量，与候选分类子集组合形成广义候选分类子集，并通过支持向量机对其筛选，确定广义基因“标签”。实验结果表明，广义基因“标签”为 4 维向量，且具有更佳的分类效果。

关键词：Bhattacharyya 距离，浮动式顺序搜索算法，RBF 支持向量机，NT_I 及 NT_Ⅱ噪声模型，LDA-GA 算法

参赛队号_____

队员姓名_____

参赛密码 _____

（由组委会填写）

一、 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。而基因在结构上发生碱基对的组成或排列顺序的改变，更改了基因原来的正常分布。因此，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 微阵列是指固定有称之为探针的核苷酸序列的固体基片或膜，它是能够快速、高效地检测 DNA 片段序列和基因表达水平的新技术。根据核苷酸分子在形成双链时所遵循的碱基互补原则，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，即基因表达谱。随着大规模基因表达谱技术的发展，已经获得人类各组织的正常的基因表达谱，为各类病人的基因表达谱提供了参考基准。如果可以在分子水平上利用基因表达谱准确地辨别是否患有肿瘤，对诊断和治疗肿瘤具有重要意义。因为正常人和肿瘤患者均具有其基因的特征表达谱，所以从 DNA 微阵列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”，能够从分子水平上准确识别是否患有肿瘤，且为医学诊断、简化实验分析及抗癌药物研制提供便捷和帮助。

然而，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，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，所以首先必须对这些“无关基因”进行剔除，然后有效地提取基因图谱信息得到基因标签。

此外，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，因此在确定肿瘤的基因标签时，应该设法充分利用其他有价值的信息，例如将与临床问题相关的主要生理学信息融合到基因分类研究中。

因此，本文需要完成以下几个问题：

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。如何根据上述观点，利用附件中的数据，选择最好的分类因素；

2． 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，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，建立融入了肿瘤研究领域中有助于诊断肿瘤信息的确定基因“标签”的数学模型。

二、 问题分析

基因表达谱作为描绘特定细胞或组织在特定状态下的基因表达种类和丰度信息，能够提供丰富的信息进行正常和患有肿瘤两类样本的辨别，为医学诊断及抗癌药物研制便捷。目前，肿瘤分类领域的一个目标是采用尽可能少的信息基因以获得尽可能高的样本分类准确率，这是因为：(1)选择尽可能少的信息基因意味着尽可能多地去掉了包含在样本中的噪音；(2)意味着减少肿瘤诊断的成本；(3)分类准确率高的信息基因通常与肿瘤的发生发展存在紧密的联系。然而，仅仅采取一种基因选择方法很难选出满足条件的信息基因子集，因此需要进行两个阶段，即初选阶段和复选阶段。初选阶段利用适当的条件限制先从成千上万个基因中选出信息基因，从而大幅降低基因的搜索空间，然后进行复选得到能有效判别正常与患有肿瘤的基因标签。

该题首先需要参赛者解决的问题是：根据 DNA 微阵列测定得到的基因表达谱，采用有效的算法，得到准确辨别正常和患有肿瘤的两类样本的基因标签，并对附件中提供的样本进行准确辨别。另外，基因表达谱中不可避免地含有噪声，会影响基因表达谱的提取，因此需要建立适当的噪声模型对基因标签筛选过程进行优化。最后，由于肿瘤是多种因素共同作用的结果，因此在确定肿瘤标签时，还要充分考虑其他有价值的信息。具体来说，需要考虑的问题如下：

1． 信息基因的初选---“无关基因”的剔除

对于某特定组织的基因表达谱，含有数量庞大的基因，其中绝大部分的基因在正常和患有肿瘤两种状态下的基因表达水平具有相似性，无法对辨别作出贡献。这类基因被称为“无关基因”。对于问题 1，首先需要选取一定标准，作为衡量某基因是否为“无关基因”的判断条件，然后对样本的基因表达谱进行筛选，剔除“无关基因”，并利用浮动顺序搜索算法得到候选分类特征子集。

2． 基因标签的选取

与患有肿瘤相关的基因数目可能含有若干个，对于问题 2，需要在问题 1 的处理结果组成的基因子集空间中，选取适当的算法，搜索得到能够准确判断正常或者患有肿瘤的基因标签。

能够使用的算法包含：支持向量机、多指标评价模型等。为了得到更为准确的基因标签，避免某次搜索受样本噪声等问题的干扰，可以进行多次搜索，每次均将支持向量机和多指标评价模型相结合进行筛选， 通过对结果的分析与评价，筛选出具有最佳分类效果的基因集合，即为基因“标签”。

3． 噪声模型的引入

对于问题 3，将噪声干扰考虑到基因表达谱的分析中，分析可能存在的各种噪声，如实验过程中的随机干扰等噪声，如果确定患有肿瘤的基因标签中某基因所占比率很小，那么在受到噪声干扰时则容易产生辨别偏差。而通过引入噪声模型排除或削弱该基因在辨别是否患有肿瘤的过程中的贡献，从而提高了分类的正确性，因此噪声模型的建立可能会对基因标签的确定产生有利的影响。

4． 在模型中融入肿瘤研究领域中有利信息

在肿瘤的研究领域内，已经存在若干有利于构建更完善的确定基因标签的信息，对于问题 4，通过完善上述数学模型，将这类信息融入到前面建立的模型中，增强基因标签判

断的准确性。通常我们会想到很多判别模型，比如：Fisher 判别法、贝叶斯判别法、支持向量机判别法等模型，在对有助于诊断肿瘤的信息具体分析后，即可尝试建立相应的判别模型。

三、 模型假设

假设一：样本中的数据真实，来源可靠，能够作为检验模型准确性的样本；
假设二：样本具有普遍性，能够作为寻找基因“标签”的依据；
假设三：样本数据里的噪声具有一般性。

四、 符号说明

符号	含义
θ	指定的 Bhattacharyya 距离的阈值
$D_{i\_\max}$	有 i 个基因的特征子集中具有最大评价函数值的基因集合
$J(D_i)$	有 i 个基因特征子集的 Bhattacharyya 距离
$K(x, x_i)$	核函数
$f_i^{(1)}$	分类准确度
$f_i^{(2)}$	被选基因数目
β	“留一法”权值
V_i	基因表达水平
$\vec{G_i}$	基因 i 的表达向量
S_V	协方差矩阵
I	染色体
τ	二进制向量
ϕ	实数向量
S_B	类间散布矩阵
S_W	类内散布矩阵

五、 模型的建立与解答

5.1 问题 1

5.1.1 理论分析

因为基因表示之间存在很强的相关性，所以对于某种特定的肿瘤，可能会有大量的基因都与该肿瘤类型识别相关。然而，在基因表达谱中，含有大量对样本类别的判别影响很小的基因。这些基因的表达水平在所有样本中都非常接近，不会为样本类型的判别提供有效的信息，反而会增加信息基因搜索的计算复杂度[1]。例如附件中给出的基因表达谱中，某些基因在健康状况正常和患有癌症两个类别里的分布，无论其均值还是方差均无明显差别，对样本类别的判定贡献很小。因此，需要剔除无关基因，缩小搜索的有效范围。

作为对基因的初选过程，需要一种适用性强、判别效率较高且容易实现的算法。因此，选择以 Bhattacharyya 距离为评价函数及浮动顺序搜索算法作为问题 1 的解决方案。

5.1.2 基于 Bhattacharyya 距离和浮动顺序搜索算法的基因分类方法

分类错误概率是模式识别中特征有效性的最佳度量，在降维空间中，特征选择的理想目标是达到分类错误概率最小，然而这点往往难于做到。因此，使得错误概率上界最小常常是一种合理的选择[7]。由 Chernoff 提出的错误概率上界是最小的，称为 Chernoff 上界。

根据 Chernoff 上界[2,3]得到误差的上边界，即：

$$P(error) \le P^{\beta}(\omega_i) P^{1-\beta}(\omega_j) \int p^{\beta}(x|\omega_i) p^{1-\beta}(x|\omega_j) dx \qquad (1)$$

其中 $0 \le \beta \le 1$，ω_i 和 ω_j 为需要判别的类别，$P(error)$ 为分类错误概率，积分部分覆盖所有特征空间，并可以等价为：

$$\int p^{\beta}(x|\omega_i) p^{1-\beta}(x|\omega_j) dx = e^{-k(\beta)} \qquad (2)$$

其中，$k(\beta) = \frac{\beta(1-\beta)}{2}(\mu_i - \mu_j)^T (\beta\Sigma_i + (1-\beta)\Sigma_j)^{-1}(\mu_i - \mu_j) + \frac{1}{2}\ln\frac{|\beta\Sigma_i + (1-\beta)\Sigma_j|}{|\Sigma_i|^{\beta}|\Sigma_j|^{1-\beta}}$，

Σ_i 和 Σ_j 为相应的协方差。

当 $\beta = 0.5$ 时，分类错误概率误差具有 Bhattacharyya 边界，并由此时 $k(\beta)$ 表达式化简得到基因的 Bhattacharyya 距离[2]，即：

$$B = k(0.5) = \frac{1}{4}\frac{(\mu_i - \mu_j)}{(\sigma_1^2 + \sigma_2^2)} + \frac{1}{2}\ln\left(\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}\right) \qquad (3)$$

上式的 Bhattacharyya 距离能够度量基因中含有的类别信息量，其由两部分组成，第一项表现了基因在两个类别中分布均值的差异对样本分类的作用；第二项体现了分布方

差的不同对样本分类的作用。这两部分具有相互促进的作用，即使基因在两类不同样本中分布的均值相同，只要分布的方差具有较大差异，仍然可以获得较大的 Bhattacharyya 距离值。而且，由式（3）可知，当某个基因的 Bhattacharyya 距离具有较大值时，$e^{-k(\beta)}$ 项具有较小值，从而分类错误概率的上界具有较小值。从模式分类[2]的角度看，某个基因的 Bhattacharyya 距离越大，表示可以利用该基因的信息进行越好的分类。

因此，利用 Bhattacharyya 距离作为衡量指标，能够较好地对样本中基因谱进行初选，剔除无关基因，得到对判别是否患有肿瘤具有帮助的信息基因集合。

附件提供的基因表达谱中，共有 62 个样本，每个样本均含有 2000 个基因的表达数据。其中，22 个样本被诊断为健康状况正常，40 个样本被诊断为患有癌症。针对两类样本，对每个基因进行 Bhattacharyya 距离计算，并作出基因的 Bhattacharyya 距离分布的直方图，如图 1 所示。

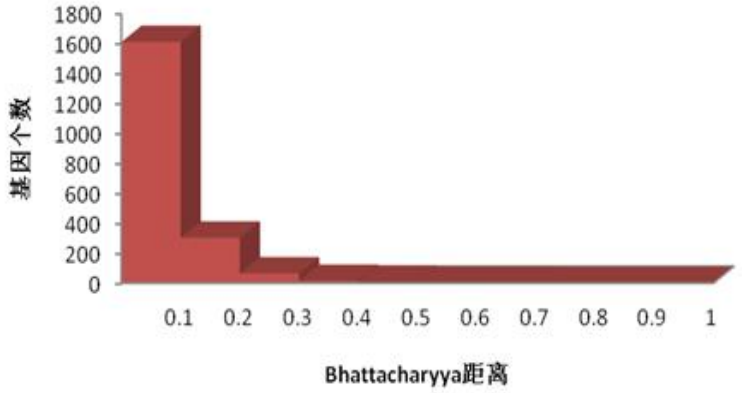

图 1 候选基因的 Bhattacharyya 距离分布直方图

根据基因所含样本类别信息的多少，选取阈值并将基因分为“信息基因”和“无关基因”两类。设 S_1 为信息基因集合，S_2 为无关基因集合，则“信息基因”与“无关基因”可以定义如下：

$$s \in \begin{cases} S_1 & B(s) > \theta \\ S_2 & B(s) \le \theta \end{cases}$$

其中 s 为基因，$B(s)$ 为基因 s 的 Bhattacharyya 距离，θ 为指定的 Bhattacharyya 距离的阈值。从图 1 可知，绝大部分基因的 Bhattacharyya 距离小于 0.1。这些基因在样本中两个类别中的分布的均值和方差均无较大差异，因此可以作为无关基因被剔除。

基因表达谱中基因 Bhattacharyya 距离的详细分布情况如表 1 所示。根据表 1 和式子可知：在阈值为 $\theta = 0.1$ 时，在 2000 个基因中，信息基因数为 388 个，无关基因数为 1612 个。其中，388 个信息基因均在一定程度上具有样本的分类信息，可以作为进一步分类的基础。

表 1 基因 Bhattacharyya 距离分布情况

Bhattacharyya 距离	基因个数	所占百分比
0~0.1	1612	80.6%
0.1~0.2	302	15.1%
0.2~0.3	63	3.15%
0.3~0.5	20	1%
0.5~1.0	3	0.15%

根据初步筛选得到的 388 个信息基因，可以形成 $2^{388} \approx 6.304 \times 10^{116}$ 个不同的基因组合，每个组合称为一个特征子集。考虑到最优搜索算法的复杂度，采用次优搜索算法，即浮动顺序搜索算法[4]对特征子集所构成的空间进行搜索，进一步得到维数不同的候选特征基因子集。

浮动顺序搜索算法(Floating Sequential Search Algorithm，FSSA)，又称增 l 减 r 算法，该算法避免了顺序前进法和后退法中特征被选入（或剔除）就无法再剔除（或选入）的缺点，在选择过程中增加了局部回溯过程[5]。

类似地，采用特征子集的 Bhattacharyya 距离 $J\left(D_i\right)$ 作为浮动顺序搜索算法的评价函数，评估特征子集对样本分类的贡献，即：

$$J\left(D_i\right)=\frac{1}{8}\left(\mu_1-\mu_2\right)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}\left(\mu_1-\mu_2\right)+\frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{\left|\Sigma_1\right|\left|\Sigma_2\right|}} \tag{4}$$

其中 $J(D_i)$ 表示具有 i 个基因特征子集 F_i 的 Bhattacharyya 距离。μ_1 和 μ_2 为特征子集 D_i 中的基因在正常和患有癌症两个类别样本中分布的均值向量，Σ_1 和 Σ_2 为对应的协方差矩阵。令 $D_{i\_\max}$ 为含有 i 个基因的特征子集中具有最大评价函数值的基因集合，它是所有维数为 i 的特征基因子集中对分类贡献最大的基因集合。

利用浮动顺序搜索算法在特征子集空间中进行搜索，得到具有不同维数的候选特征子集 $D_{i\_\max}$。$FFSA\left(n,\left\{D_{i\_\max}\middle|\dim\left(D\right)_i=i,i=1,2,\cdots,n\right\}\right)$ 算法具体步骤如下：

step1：初始化 i=2，n=50，$D_{2\_\max}=\{g_1,g_2\}$ g_1,g_2 为 388 个候选基因集 G_0 中 Bhattacharyya 距离最大的两个基因；

step2: $G=G_0-D_{2\_\max}$，G 即为候选基因集中去掉当前基因子集的其余基因组成的几何；

step3:建立新子集 $D_{(i+1)\_\max}=\{D_{i\_\max}+g\}$,其中 $g\in G$ 并且 $J(D_{(i+1)\_\max})=\max J(D)$ ，$D\in\{D_{i+1}\mid \dim(D_{i+1})=i+1, D_{i+1}\subset G_0\}$；

step4：搜索新的子集 $G'_{i\_\max}\subset G_{(i+1)\_\max}$，使 $G'_{i\_\max}=\arg(\max(J(D')))$，其中 $D'\in\{G_i\mid \dim(G_i)=i, G_i\subset G_{(i+1)\_\max}\}$；

step5：若 $J(D'_{i\_\max})\le J(D_{i\_\max})$，则 $i=i+1$，转 step7；否则令 $D_{i\_\max}=D'_{i\_\max}$ ；

step6：如果 $i=2$，转到 step2；否则，$i=i-1$，转到 step4 ；

step7：如果 $i=n$ 或 Bhattacharyya 距离评价函数中开始出现奇异协方差矩阵，退出；否则，转 step2。

浮动顺序搜索算法的算法流程图如图 2 所示。

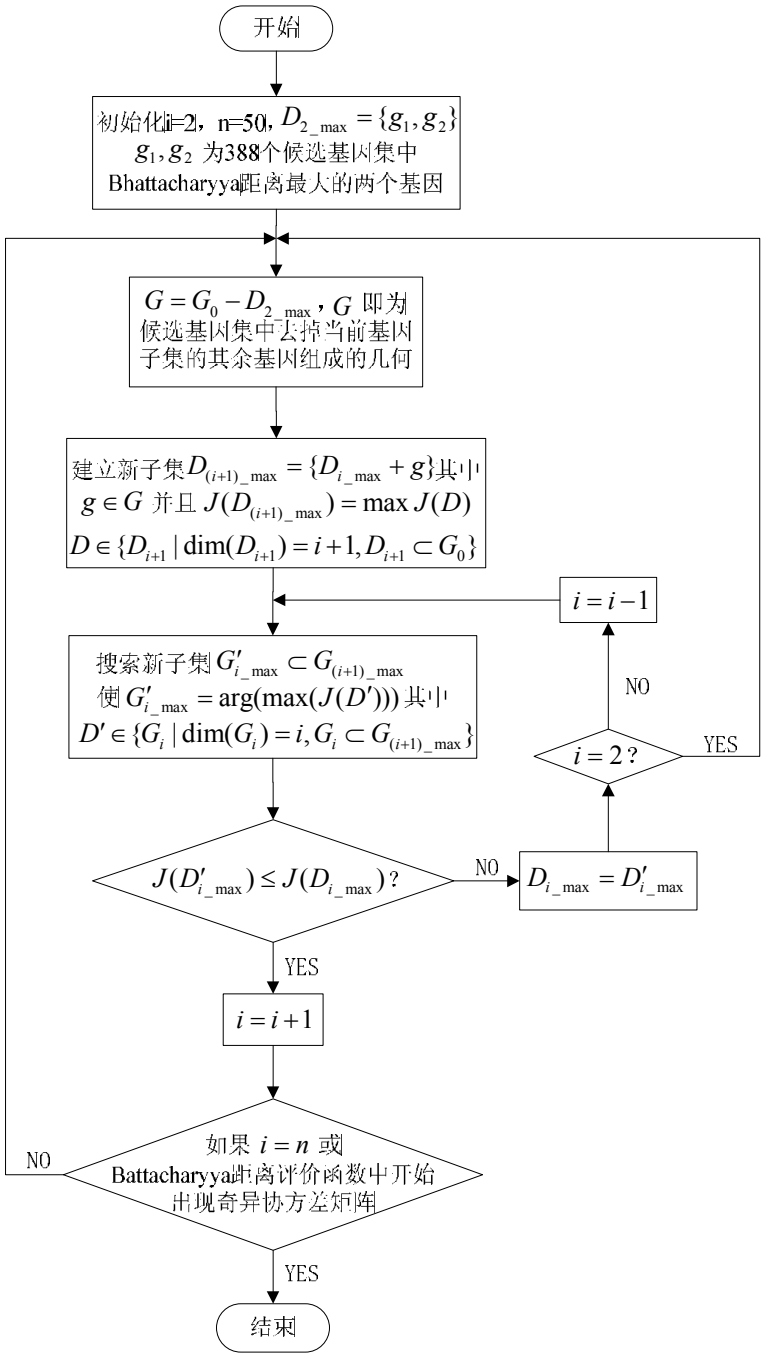

图 2 浮动顺序搜索算法流程图

附件样本中有正常样本 22 个，肿瘤样本 40 个，在执行浮动顺序搜索算法过程中，式（3）中的协方差矩阵$\sum_1$出现奇异，根据该算法中 step7 的截止条件，程序运行结束，此时$i=36$，因此候选特征基因子集的最大维数为 36，并得到 35 个具有维数不同的候选基因特征子集$D_{i_\max}, i=2,3,\cdots,36$，如附表 1 所示。

5.1.3 小结

对于第 1 问，首先利用 Bhattacharyya 距离作为评价指标进行基因的初选，剔除无关基因，从样本中的 2000 个基因中得到 388 个信息基因，之后在信息基因所生成的特征基因子集空间内，利用浮动顺序搜索算法，得到 35 个具有维数不同的候选特征基因子集，每个候选特征基因子集均是在具有对应维数的特征基因子集中评价函数值最大的。

从结果上讲，所选取的 Bhattacharyya 距离评价函数及浮动顺序搜索算法适用于题中要求，能够高效地对基因表达谱中的无关基因进行剔除，因此可以作为最好的分类因素。所得到的结果可以作为第 2 问中基因标签选取的基础。

5.2 问题 2

5.2.1 理论分析

相对于基因数目，样本的数量很小，如果直接用于分类会造成小样本的学习问题。另外，分类准确率高的信息基因通常与肿瘤的发生发展具有密切的联系，因此，为了得到较好的分类效果，需要减少用于分类识别的信息基因数目。

支持向量机（SVM）[8]是由 Vapnik 领导的 AT&T Bell 实验室研究小组提出的一种非常有潜力的新的分类技术。支持向量机建立在统计学习理论的 VC 维理论，并利用结构风险最小化原则，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷。此外，支持向量机算法是一个凸二次优化问题，能够保证找到的极值解同时也是最优解，而且，目前有许多优化算法可以用来解决此类凸优化问题，为支持向量机的可行性提供了保障。因此，支持向量机是具有很强的理论依据和理论基础并具有很强的应用价值的建模方法。

支持向量机（SVM）在解决小样本、非线性及高维模式识别中具有许多优势，其目标是针对小样本情况，得到现有信息下的最优解，具有坚实的数学和理论基础。此外，支持向量机最终得到的解是全局最优解，解决了在神经网络方法中无法避免的局部极值问题，而且其算法复杂度与样本维数无关，非常适合处理非线性问题，并具有非常好的推广能力。根据问题 1 所得结果，需要处理的样本的数目为 62 个，维数为 36 维，而且需要解决的问题是将样本分为正常和患有肿瘤两种类别。因此从分类角度来讲，支持向量机对于解决本题中的分类问题具有很强的理论基础和可行性。

5.2.2 基于支持向量机的基因标签选择

5.2.2.1 支持向量机

支持向量机的基本思想是，首先把训练数据集非线性地映射到一个高维特征空间（即 Hilbert 空间），这个非线性映射的目的是把在输入空间中的线性不可分数据集映射到高维特征空间后变为线性可分的数据集，随后在特征空间建立一个具有最大隔离距离的最优分离超平面，这也相当于在输入空间产生一个最优非线性决策边界[9]。如图 3 所示。

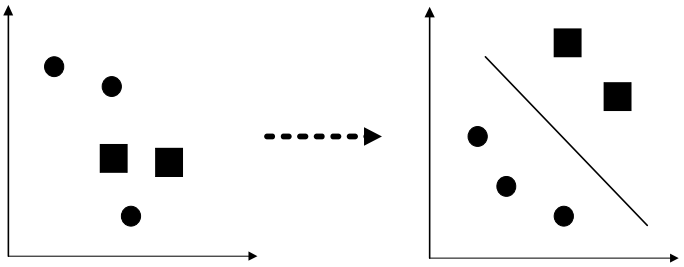

图 3 支持向量机的基本思想

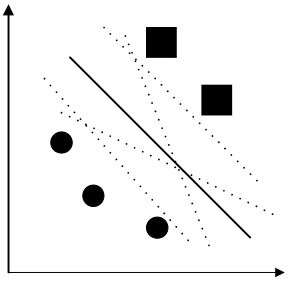

图 4 最优分离超平面和非最优分离超平面

在特征空间中支持向量机的分离超平面是最优的分离超平面，其最优性可以从图 4 中看到。图中存在多个分离超平面均可将两个类分离开，然而只有图中实线所示超平面为最优超平面，因为其与两个类之间的最近向量距离最大。从几何上讲支持向量就是决定最优分离超平面的样本向量的最小个数。

在构造支持向量机时必须解决两个关键问题：其一是找到一个非线性映射，将线性不可分数据集映射到高维特征空间中的线性可分数据集；其二是在高位特征空间中求取最优分离超平面。

对于第一个问题，可以利用核技术和方法解决。常用的核具有以下几种：

(1)线性核： $K\left(x,x_i\right)=x_i\cdot x$

(2)径向核： $K\left(x,x_i\right)=\exp\left(-\left\|x-x_i\right\|^2\Big/2\sigma^2\right)$

(3)多项式核： $K\left(x,x_i\right)=\left(x_i\cdot x+1\right)^d, d=1,2,\cdots,N$

很多的应用和经验表明，径向基函数（RBF）支持向量机具有良好的学习能力。

对于第二个问题，可以证明其为一个典型的受约束二次型规划问题。用作二分类时支持向量机的输出为：

$$f\left(x,W\right)=\mathrm{sgn}\left(\sum_{i=1}^{N}\omega_i K\left(x,x_i\right)+b\right) \tag{5}$$

优化目标函数为： $J=W^TW=\left\|W\right\|^2$ （6）

约束条件为： $y_i\left[\sum_{i=1}^{N}\omega_i K\left(x_j,x_i\right)+b\right]\geq 1, j=1,\cdots,N$

其中 N 是样本数目，W 是支持向量机的输出可调参数向量，$\left(x_i,y_i\right)$ 是样本。目标函数是为了保证分类的最优性，约束条件是为了保证分类的正确性。为了消除噪音和异常样本的影响，引入松弛变量如下：

$$\begin{aligned} &J=\frac{1}{2}W^TW+C\sum_{i=1}^{N}\xi_i \\ &y_i\left[\sum_{i=1}^{N}\omega_i K\left(x_j,x_i\right)+b\right]\geq 1-\xi_i, j=1,\cdots,N \\ &\xi_i\geq 0, j=1,\cdots,N \end{aligned} \tag{7}$$

5.2.2.2 基于支持向量机的肿瘤分类样本的输入与输出

（1）输入

通过前面浮动搜索算法，本文共得到 35 个候选分类特征子集。每个特征子集都需要通过支持向量机来检验其分类效果。则第 i 个分类特征子集所对应的支持向量机分类问

题的输入为：$X_i=\{x_1,x_2,...,x_{i+1}\}$，其中$i=1,2,...,35$，每个分类特征子集所含的元素详见表 1。

（2）输出

样本分为两种类型：一种为 Normal 类，用-1 表示；另一种为 Cancer 类，用 1 表示。对于第i个分类特征子集，若它属于 Normal 类，则$y_i=-1$；否则，$y_i=1$。

5.2.2.3 基于支持向量机的基因标签选择过程

由于样本数量较少，为了获得对候选特征基因子集分类错误率较为可靠的估计，需要在训练集和测试集上分别做分类错误率估计，样本数据集中共有 62 组数据，其中 22 组数据为 Normal 类型，其余 40 组数据为 Cancer 类型。为了将样本分为训练集和测试集，随机选出 40 个样本作为训练集，其中 14 个样本为 Normal 类型，其余 26 个样本为 Cancer 类型；剩下的 22 个样本作为测试集，其中 8 个样本为 Normal 类型，其余 14 个样本为 Cancer 类型。

(1) 在训练集上采用“留一法”（LOOCV）进行样本类别的辨别。训练集中共有 40 个样本，每次保留 1 个样本作为测试样本，其余 39 个样本作为支持向量机的训练样本。循环 40 次，使得每个样本均能用作测试样本。累计被错误分类的样本数，作为“留一法”分类错误数。留一法的基因标签分类效果检验过程如图 5 所示：

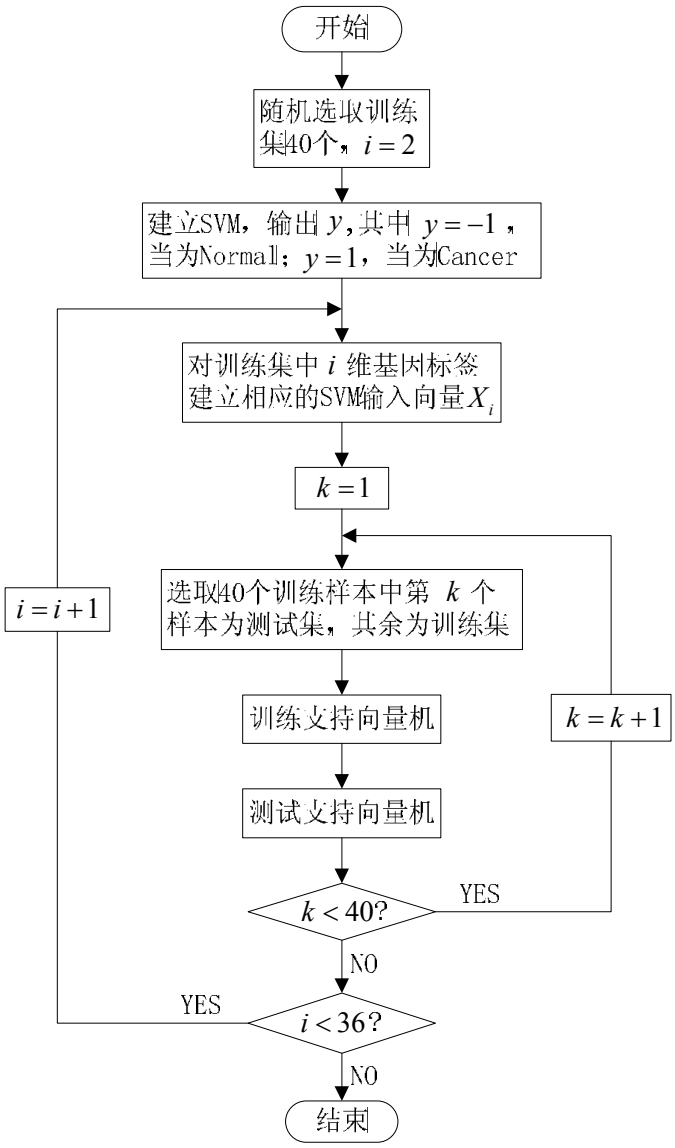

图 5 基于“留一法”的基因标签分类效果检验过程

(2) 用训练集中所有的 40 个样本训练支持向量机，然后识别测试集中 22 个样本的类型，被错误分类的样本数为“独立测试实验”的分类错误数，其之分类效果如图 6 所示：

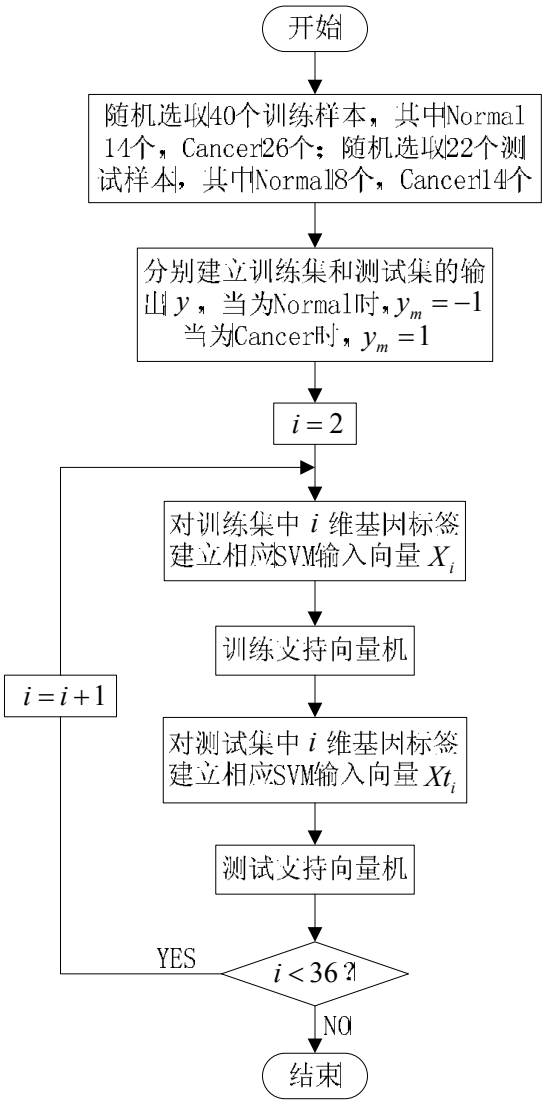

图 6 基于“独立测试实验”的基因标签分类效果检验过程

5.2.3 基于“多指标评价模型”的筛选结果

基于支持向量机分别采用“留一法”和“独立测试实验”检验候选特征子集的分类效果，由上面的测试过程介绍及流程图可知，“留一法”的测试次数为 40 次，“独立测试实验”的测试次数为 22 次。先将它们的测试效果以误分次数的方式显示于图 7 中：

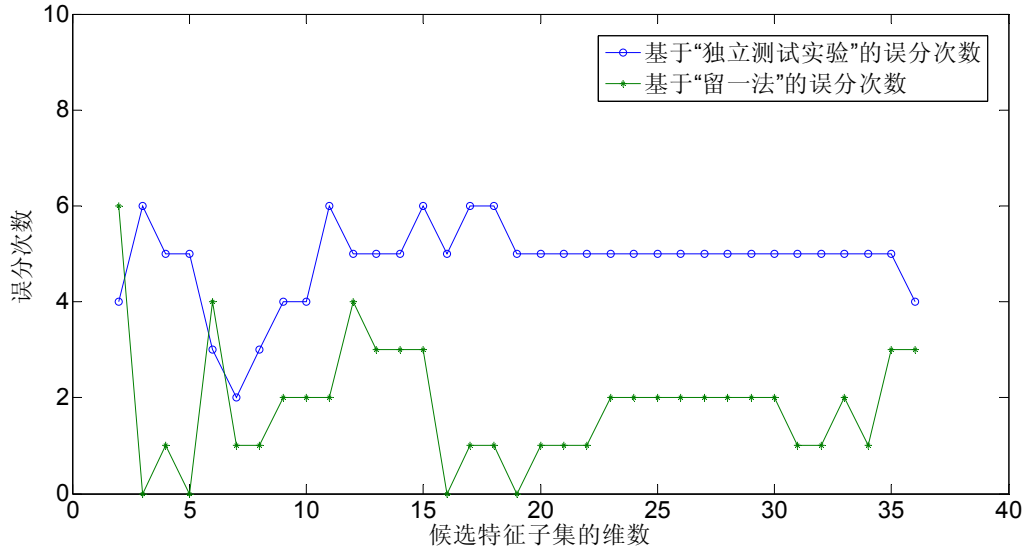

图 7“留一法”以及“独立测试实验”误分次数

由图 7 可知，“留一法”和“独立测试实验”所得的结果均可以用于衡量候选分类特征子集的分类效果，另外，候选分类特征子集的维数也是衡量其优劣的重要指标，因此，候选特征子集的衡量标准由以上三个，下面利用一种有效的评价算法筛选处最优的候选特征子集。

通过分类准确度（ $f_i^{(1)}$ ）和被选基因数目（ $f_i^{(2)}$ ）来衡量一个候选分类特征子集优越性。

第 i 个候选特征子集的评价函数 f_i 定义为下述的加权求和：

$$f_i = (1-\alpha)\cdot f_i^{(1)} + \alpha \cdot f_i^{(2)} \qquad 0 \le \alpha \le 1 \tag{8}$$

其中， α 是一个参数，其允许我们给 $f_i^{(1)}$ 和 $f_i^{(2)}$ 分配一个相对重要因素。给 α 赋一个大于 0.5 的值，评价函数将会侧重于高分类精度（可能会以选择更多的基因为代价）。相反，使用小的 α 值将会侧重于小的基因子集。因此，变化的 α 会改变评价函数的评价方向。

值得注意的是 f_i 的取值范围是 $[0,1]$；那么，一个 f_i 值大的解就比一个 f_i 值小的解好。

第 i 个候选特征子集的精度评价函数 $f_i^{(1)}$ 定义如下：

$$f_i^{(1)} = \beta \cdot f_i^{(11)} + (1-\beta)\cdot f_i^{(12)} \tag{9}$$

β 为“留一法”与“独立测试实验”权值。

第 i 个候选特征子集的维度评价函数 $f_i^{(2)}$ 的定义如下：

$$f_i^{(2)} = 1 - (i+1)/D \tag{10}$$

$i+1$ 为选择基因个数， D 为最大维数。

第 i 个候选特征子集基于“留一法”测试的评价函数的定义如下：

$$f_i^{(11)} = 1 - error_i^{LOOCV} / N_{LOOCV} \tag{11}$$

$error_i^{LOOCV}$ 第 i 个候选特征子集在“留一法”中的错分次数， N_{LOOCV} 为“留一法”样本数量。

第 i 个候选特征子集基于“独立测试实验”测试的评价函数的定义如下：

$$f_i^{(12)} = 1 - error_i^{TEST} / N_{TEST} \tag{12}$$

$error_i^{TEST}$ 第 i 个候选特征子集在“独立测试实验”中的错分次数， N_{TEST} 为“留一法”样本数量。

由式（a1）、式（a2）和式（a3）知， $f_i^{(1)}$ 越大，该特征子集的精度越高；由式（a4）可知， $f_i^{(2)}$ 越大，该特征子集的维数越小；因此，最后由式（a5）可知， f_i 的值越大，该特征子集越优秀。

由于 α 表示了维度与精度之间的权衡关系，因此其取值较难确定，且对筛选结果具有重要影响，为了准确的评价各个特征子集，需要使改评价对于 α 具有较强的容忍度，图 XXX 即为当 α 在 0.3 到 0.7 之间连续变化时评价值的变化。

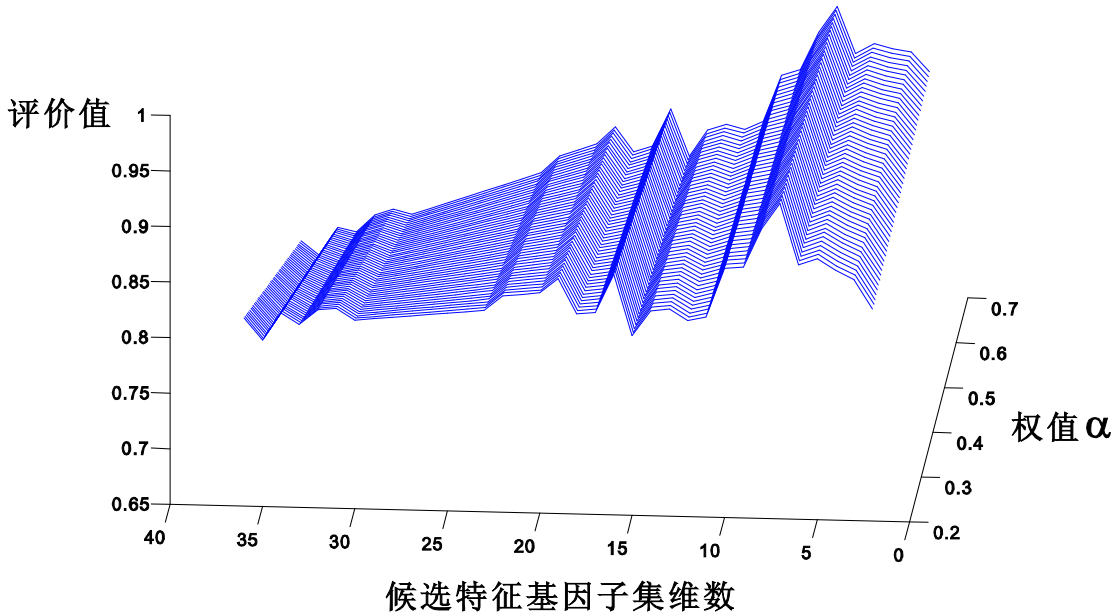

图 8 基于多指标评价函数的候选特征基因子集评价

由图 XXX 可知，无论α取 0.3 到 0.7 之间的任意值，维数为 7 的特征子集始终为具有最高评价值的特征子集，因此，该特征子集即为具有最佳分类效果的特征子集，即该特征子集为基因“标签”。查表“附表 1”可知，该特征子集以序号表示为：{22， 698，1897，1346，1924，1740，1352}。查找数据集，其进一步的信息如表 XX 所示：

表 2 基因“标签”介绍

表达序列标签	基因序列数据库标号	基因描述
Hsa. 6080	J02763	Human calcyclin gene, complete cds.
Hsa. 9972	T51261	GLIA DERIVED NEXIN PRECURSOR (Mus musculus)
Hsa.466	U19969	Human two-handed zinc finger protein ZEB mRNA, partial cds.
Hsa. 5392	T62947	60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)
Hsa. 43331	H64807	PLACENTAL FOLATE TRANSPORTER (Homo sapiens)
Hsa. 2012	M81651	Human semenogelin II (SEMGII) gene, complete cds.
Hsa. 712	M24069	Human DNA-binding protein A (dbpA) gene, 3' end.

5.3 问题 3

5.3.1 噪声模型描述

在 DNA 芯片所测量的成千上万个基因中，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，在肿瘤分型中这类基因应该被剔除。我们将这类基因称之为Ⅰ类噪声基因，简称 NT-Ⅰ基因。

在本文中着重处理这样一类基因：它们在大多数样本中表达水平非常接近，只有在极少数样本中出现了特殊的表达状况。这些基因也可能是 NT-Ⅰ基因，由于实验过程中的随机干扰导致某些实验数据发生了很大的变化，这些数据不能真实反映基因的表达水平，在肿瘤分型中应该被剔除。

还有一类基因：这些基因确实反映了样本的特殊性，即使在一个肿瘤诊断分类中，每个癌症患者都具有一个特殊的基因表达谱，具有独一无二的肿瘤类型。但是，从目前的测量技术手段获得的数据质量来看，还无法达到将每个个体作为一个亚型来处理的水平。也就是说，我们希望用比较细的尺度来诊断肿瘤类型，但是当这个尺度细化到一定水平，信号就被噪声淹没了。如果将这些基因包括在后续的肿瘤分类过程中，必将提高噪声，将多个样本中的共性模糊化。若将这类基因过滤掉以后，对肿瘤分类就可以得到比较准确的结果。我们将这些基因称之为Ⅱ类噪声基因，简称 NT-Ⅱ基因。

为了剔除 NT-Ⅰ基因，将每个基因在所有样本中的变化情况和一个预设定阈值 L 进行比较。基因表达水平的变化定义如下：

$$V_i = \max\left(\vec{G}_i\right) \Big/ \min\left(\vec{G}_i\right) \tag{13}$$

式中，$\vec{G}_i$ 指的是基因 i 的表达向量，V_i 表示 $\vec{G}_i$ 的变化值。如果 $V_i < L$，就认为基因 i 是一个 NT-Ⅰ基因，将被剔除掉。运用这个准则对所有的基因进行过滤，得到的基因集合记为 X。

采用一种重采样技术对 NT-Ⅱ基因进行剔除，输入是 X，首先从所有样本中随机抽取一个子集，然后使用Ⅰ型噪声基因的过滤方法对该子集进行过滤。得到的基因集合记为 Y，然后令 $X = Y$，依次迭代此过程。

对每次迭代获得的 Y 中的基因数量，画一条曲线来表示 Y 中的基因数量的变化情况，发现该曲线在开始阶段很快下降，因为大部分的Ⅱ型噪声基因在重采样的开始就被过滤掉了。随着Ⅱ类噪声基因的减少，该曲线将变得平滑，我们选择最小曲率半径附近的点作为最优迭代步骤，并确定相应的基因集合。

5.3.2 基因表达谱中噪声的排除

（1）NT_I 类噪声的排除

① 利用式（13）计算所有的基因的变化值，如图 9 所示：

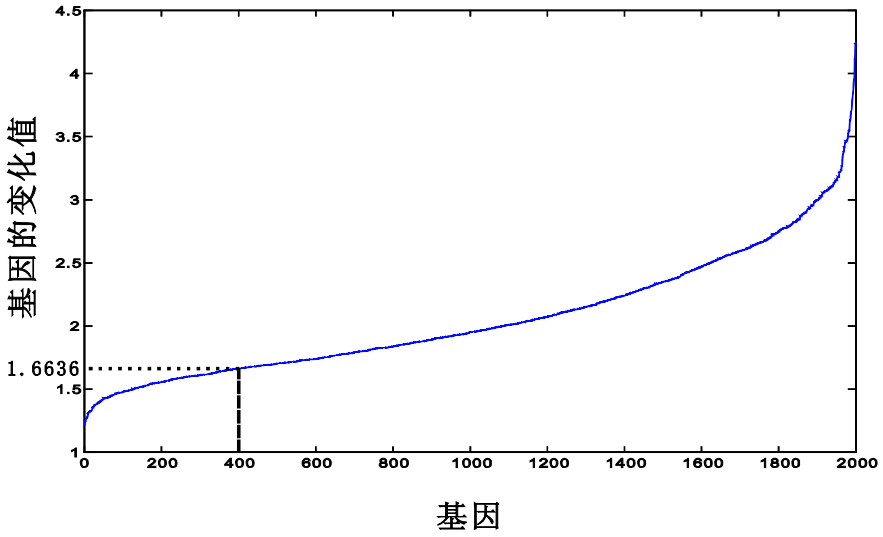

图 9 通过判断基因的变化值来筛选 NT_I 类噪声

②由图 9 可知，在第 400 个基因点附近时，该曲线具有最小的坡度，选取阈值 $L=1.6636$，则有 400 个基因点由于变化值太小而被作为 NT_I 类噪声基因排除，基因数减小为 1600 个。

（2）NT_II 类噪声基因的排除。

①利用前述基于重采样技术的迭代算法不断排除 NT_II 类噪声基因，将每次迭代所剩的基因数绘制成为图 10：

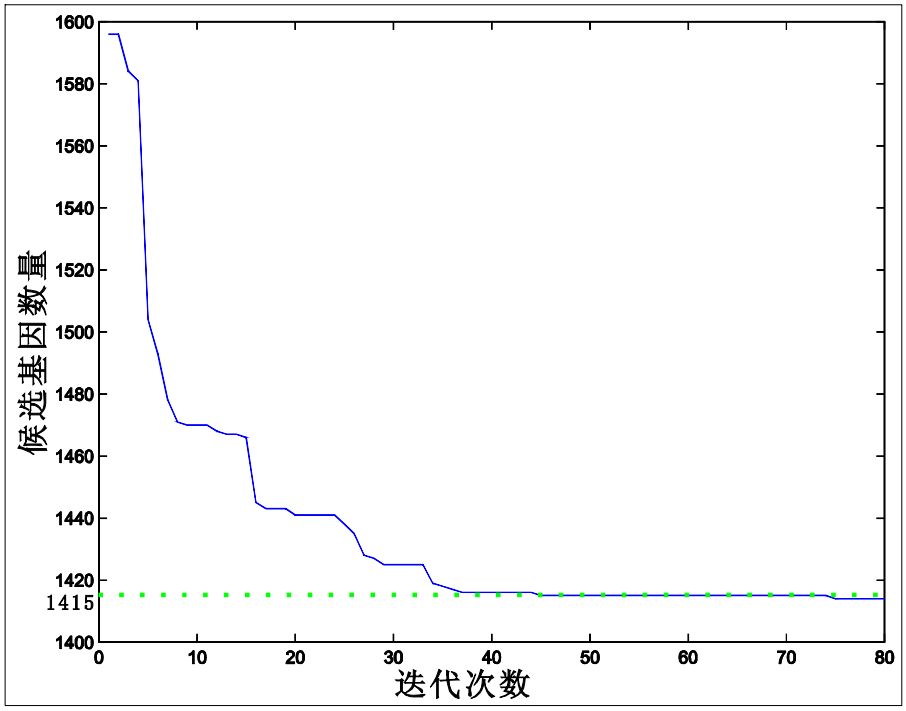

图 10 候选基因迭代数量随迭代次数的变化

②根据图 10，选择图中平缓部分对应的基因数量为排除 NT_II 类噪声基因后的基因个数，因此剩余基因数目有图选为 1415 个，这 1415 个基因组成的集合即为排除 NT_I 以及 NT_II 类噪声基因后的基因集合。

5.3.3 排除噪声后的效果检验

（1）计算去噪后的基因集合中每个基因的 Bhattacharyya 距离，绘制出 Bhattacharyya 距离分布直方图：

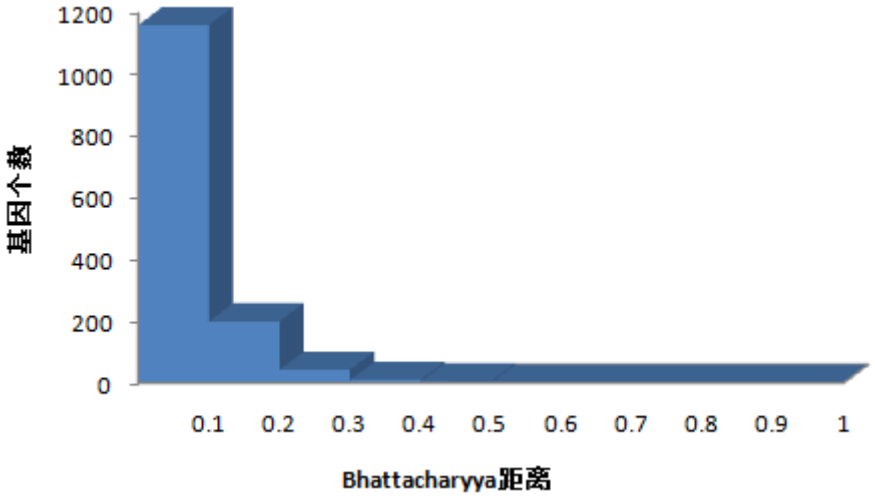

图 11 候选基因的 Bhattacharyya 距离分布直方图

（2）由图 11，筛除 Bhattacharyya 距离小于 0.1 的基因，剩下的基因总数为 254。

（3）执行浮动搜索算法，得到 35 个候选特征分类子集，其详细情况见附图 2。

（4） 基于支持向量机对候选分类特征子集进行效果测试，其中“留一法”的测试次数为 40 次，“独立测试实验”的测试次数为 22 次。先将它们的测试效果以误分次数的方式显示。图 12 中 35 个候选特征分类子集分别训练支持向量机，经“留一法”和“独立测试实验”得到其分类效果，如图 12 所示：

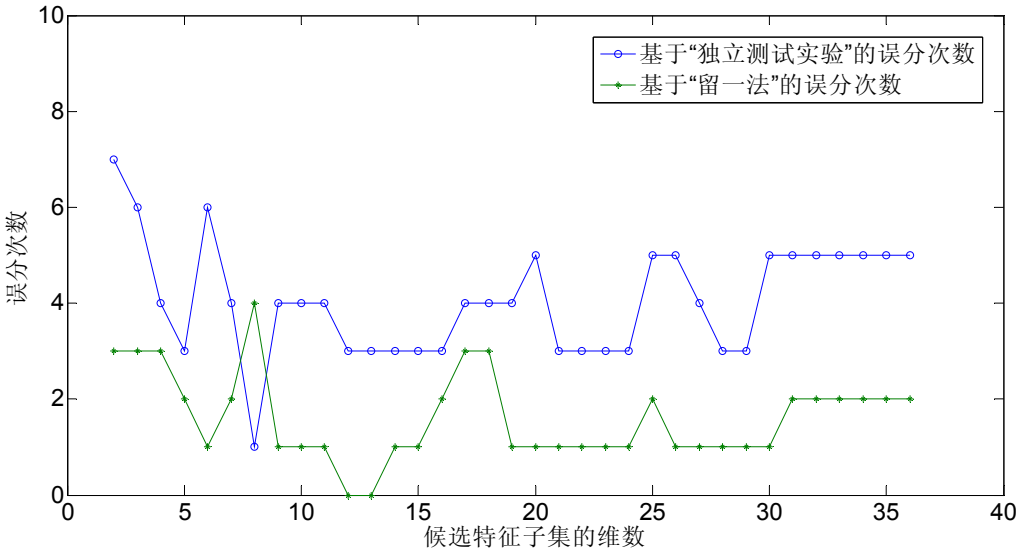

图 12“留一法”以及“独立测试实验”误分次数

根据图 12，再由式（8）、式（9）、式（10）、式（11）和式（12）知，f_i的值越大，该特征子集越优秀，而最优的特征分类子集即为基因“标签”。

（5）为了准确的评价各个特征子集，需要使改评价对于α具有较强的容忍度，图 13 即为当α在 0.3 到 0.7 之间连续变化时评价值的变化。

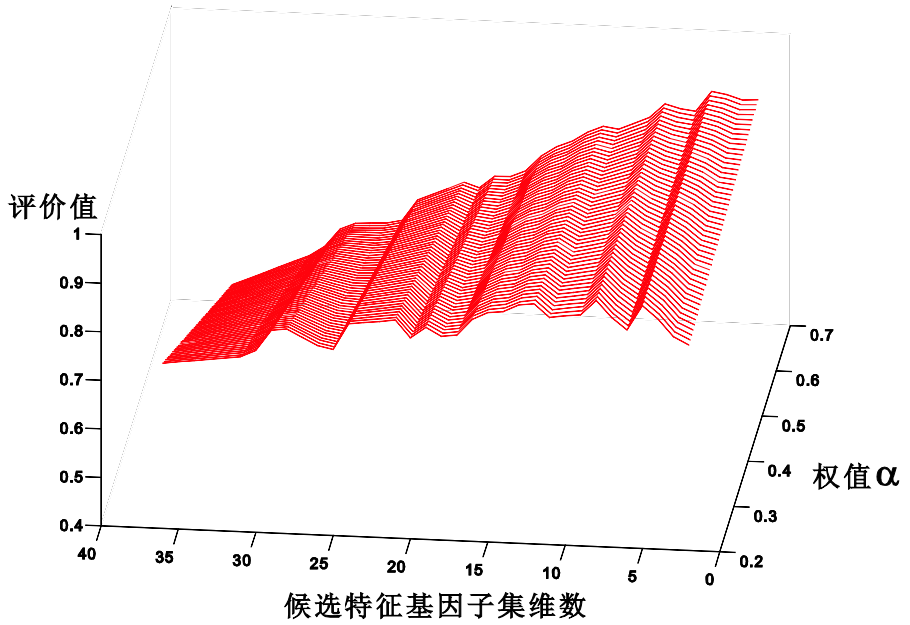

图 13 基于多指标评价函数的候选特征基因子集评价

由图 13 可知，无论α取 0.3 到 0.7 之间的任意值，维数为 5 的特征子集始终为具有最高评价值的特征子集，因此，该特征子集即为具有最佳分类效果的特征子集，即该特征子集为基因“标签”。查表“附表 2”可知，该特征子集以序号表示为：{ 1582，493，1346，586，1870}。

表 3：基因“标签”介绍

表达序列标签	基因序列数据库标号	基因描述
Hsa. 2928	X63629	H.sapiens mRNA for p cadherin.
Hsa. 37937	R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
Hsa. 5392	T62947	60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)
Hsa. 608	H17897	ADP,ATP CARRIER PROTEIN, FIBROBLAST ISOFORM (HUMAN);.
Hsa. 1660	H55916	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL

5.3.4 去除噪声对确定基因“标签”的有利影响

（1）如图 14 所示，其中虚线表示的数据点集为去噪之前的误分次数，实线表示的是去噪之后的误分次数，可以看出无论是“独立测试实验”还是“留一法”来评价误分次数，去噪之后的误分次数的平均水平普遍低于去早之前，尤其是“独立测试实验”中，去噪后的误分次数比较明显的低于去噪之前。

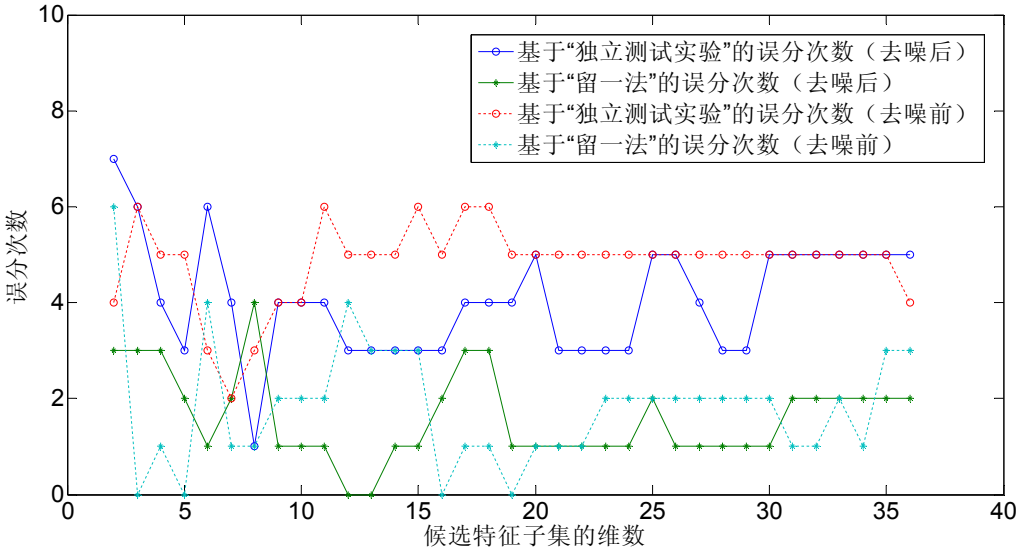

图 14 去噪前后“留一法”以及“独立测试实验”误分次数

（2）如图 15 所示，其中红色曲面表示去噪之后的候选特征基因子集的评价值，蓝色表示的是去噪之前的评价值，可以看出，红色的曲面基本高于蓝色曲面，说明经过去噪之后候选特征基因子集的分类效果普遍高于去噪之前。而且两个曲面的最高点属于红色曲面，说明去噪之后的最优候选特征基因自己（即基因“标签”）优于去噪之前，这意味着①去噪之后所筛选出的基因标签叫去噪之前有更好的分类效果，②而且去噪之后的基因“标签”的维数低于去噪之前。总之，去除噪声对确定基因“标签”的具有有利影响。

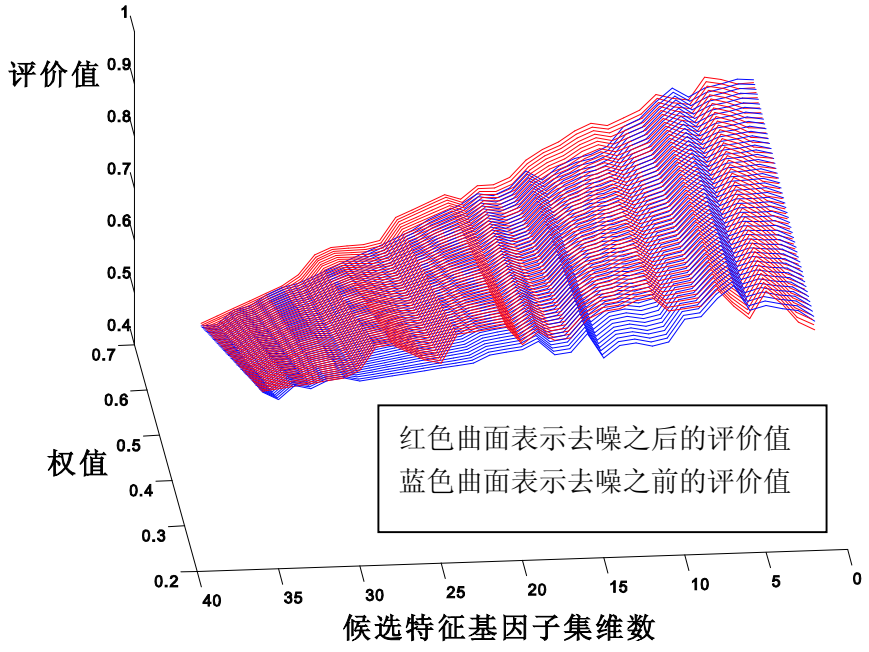

图 15 去噪前后基于多指标评价函数的候选特征基因子集评价

5.4 问题 4

通常情况下，肿瘤研究领域中会已知若干个生理基因与某种癌症的关系紧密，而有些专家[10]指出在基因分类研究中不应该忽略基因低水平表达、差异不大表达的情况，并应当将临床问题相关的主要生理学信息融入到基因分类的研究中。最后，根据以上信息建立融入了有助于诊断肿瘤信息的确定基因“标签”的模型。

5.4.1 理论分析

肿瘤研究领域中有若干重要信息，对判别是正常或者患有肿瘤具有重要的贡献，比如题中提供的临床生理学信息：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。

基因失活在本题中是指基因不表达或者表达值降低。因此，在选择基因“标签”的过程中，不仅要从基因突变导致癌变的角度考虑，同时可以通过引入基因失活等特征对模型进行优化。从样本中的基因表达谱中找到 APC 基因，其基因表达值均值和方差均较小，参考前三个问题中针对基因“标签”的求取方法，可见 APC 基因在基因标签的选择中被过早的剔除。然而，根据临床的生理学信息，这 APC 基因对判定是否患有肿瘤具有较大的判定作用，因此在筛选的过程中，需要保证 APC 等在肿瘤研究领域对分类具有重要作用的基因被保留，从而增强模型的判别准确性。

因此，提出广义基因“标签”的概念，即在原始基因标签基础上增加临床上具有重要分类信息的基因。具体的广义基因“标签”的选择方法是，在肿瘤研究领域中具有重要判别信息的基因中，首先利用 LDA-GA 方法进行筛选，得到一定维数的最优生理基因向量，然后将其加入到去噪后的每一个候选分类特征子集，得到广义候选分类特征子集，然后利用 RBF 支持向量机进行筛选，得到具有最佳分类效果的广义基因“标签”。具体过程如图 16 所示。

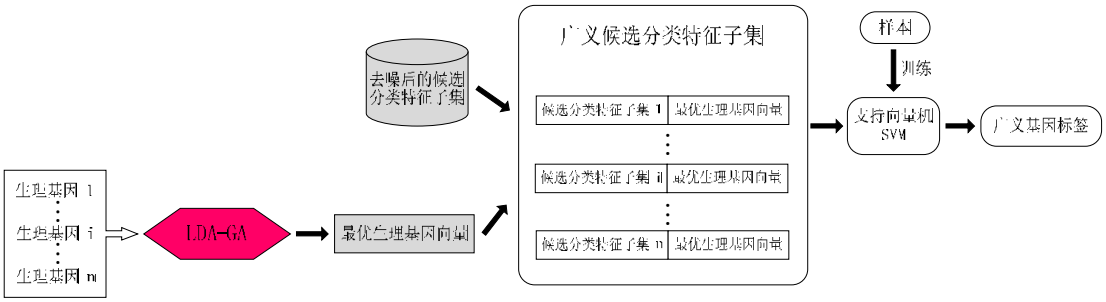

图 16 广义基因“标签”确定模型框图

5.4.2 建立确定广义基因“标签”的数学模型

临床上，有很多对判别是否患有肿瘤具有非常重要贡献的依据，这些信息涉及的范围较广，如果仅仅从信息基因的角度进行辨别，会出现判别不准确的情况。然而，针对肿瘤这种特殊重症的诊断需要很高的准确性。因此，需要从多方面对此分类问题进行考察，尽可能地扩大考察范围，得到较为准确的数学模型。

根据文献中的数据，BRCAI 抑癌基因定位于 17 号染色体长臂，该基因的突变引起恶性肿瘤发生的易感器官有乳腺、卵巢、结肠及前列腺等，统计表明乳腺癌-卵巢有 80%～90% 伴有 BRCAI 突变[11]；P16 蛋白多重肿瘤抑制基因（MTSI）的表达产物，MTSI 定位于人类第 9 号染色体短臂不到 40*kb* 范围内，MTSI 基因在肿瘤中的总突变率为 75% [12]；PTEN 基因定位于染色体10*q*23.3，含有 9 个外显子，定位于 10 号染色体。该

基因在许多进展期的肿瘤中均有发现。PTEN 基因在大肠癌中的突变率为 75% [13]。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[1]。以上数据能够作为具有代表性的生理基因参与筛选，筛选过程具有很强的典型性。

在和癌症相关的生理信息基因集合中，为了选择一个小的生理信息基因子集，本文采用一个嵌入式算法来实现这个目标，这个基于 LDA 的遗传算法将遗传算法和 Fisher 线性判别分析相结合，其主要优势是：不仅将 LDA 分类器融入遗传算法的适应度函数中，而且在遗传算法的交叉和变异算子中也融入了 LDA 的判别系数。这样就把与问题相关的信息融入到选择操作中了，因此使用该算法可以实现高精度的肿瘤判别。

首先对 LDA-GA 方法进行简单介绍。

（1） 线性判别算法（LDA）

LDA 是众所周知的降维和分类算法，它是通过将数据投影到一个低维空间来进行数据的较佳分类。近年来，LDA 更多地被应用在微阵列数据的分析中。

在利用 LDA 方法解决二分类问题时，首先需要明确这个问题的定义。考虑一个含有 C_1 和 C_2 两个类别的样本集合，其中 C_1 类含有 n_1 个样本数据，C_2 类含有 n_2 个样本数据，而每个样本均能用 q 个变量来描述。因此，对于矩阵 $X=\left(x_{ij}\right), i=1,\cdots,n;\ j=1,\cdots,q$ 中的元素，我们定义 μ_k 为集合 C_k 的均值，μ 为所有样本的均值，即：

$$\mu_k=\frac{1}{n}\sum_{x_i\in C_k}x_i \tag{14}$$

$$\mu=\frac{1}{n}\sum_{x_i}x_i=\frac{1}{n}\sum_{k}n_k\mu_k \tag{15}$$

数据可以利用 S_B 和 S_W 两个矩阵描述，其中 S_B 为类间散布矩阵，S_W 为类内散布矩阵。

$$S_B=\sum_k n_k\left(\mu_k-\mu\right)\left(\mu_k-\mu\right)^T \tag{16}$$

$$S_W=\sum_k\sum_{x_i\in C_k}\left(x_i-\mu_k\right)\left(x_i-\mu_k\right)^T \tag{17}$$

如果定义 S_V 为协方差矩阵，那么 $S_V=S_B+S_W$。

LDA 算法的目标是寻找一个初始变量的线性组合，使得两类的均值较好地分离，其能通过被分配到每个类别中的数据变量的总和来量度。因此，LDA 最终确定向量 w，使得当 w^TS_Ww 具有最小值时，w^TS_Bw 有最大值。通过向量 w_{opt} 来实现这个双目标优化，可以通过最大化下式来得到 w_{opt}：

$$J(W)=\frac{w^T S_B w}{w^T S_W w} \tag{18}$$

可以证明，当 S_W^{-1} 存在时，向量 w_{opt} 是 $S_W^{-1}S_B$ 的单个特征值所对应的特征向量。一旦 w_{opt} 的基准线被确定，LDA 就会提供一种分类方法（分类器）。然而，在本文中，我们特别关注这个向量的判别系数：这些系数的绝对值表示 q 个初始变量对分类的重要性。

（2） 基于 LDA 的遗传算法（LDA-GA）

利用基因滤波算法对生理信息基因进行初选后，对于一个含有 p 个基因的集合，基于 LDA 的遗传算法被用来执行在大小为 2^p 的空间中的组合搜索。搜索的目的是在可能的基因组合中确定具有高预测精度的最小基因子集。下文中将会提出该算法的一般步骤，并对基于 LDA 的遗传算法中的成分进行分析。特别地，将会解释 LDA 和遗传算法是如何进行结合的。

①┌ 一般遗传步骤

本文的基于 LDA 的遗传算法遵循如下传统遗传算法的模板，并具有优秀的策略。

- 初始种群：初始种群是在每条染色体所包含的基因数目范围从 $p\times0.6$ 到 $p\times0.75$ 之间的集合中随机生成的。
- 进化：当前种群 P 的染色体是根据适用度函数进行排序的。种群 P 中染色体最“优”的10% 被直接复制到下一个种群 P'，并从 P 中移除。P' 中剩下 90% 的染色体由交叉和变异生成。
- 交叉和变异：父染色体是在 P 中剩下的染色体中，根据相邻染色体对决定的。利用特别的交叉算子，每次产生一个子体。这个子体在加入下一个种群 P' 前经历一个变异过程。
- 终止条件：当到达一个预先设定好的迭代数目，或者当种群中出现一条染色体的基因子集很小时，进化过程则终止。

② 染色体编码

通常地，一条染色体简单地用来代表一个候选基因子集。这里提到的遗传算法中，一条染色体具有更多的信息，并由一对向量定义如下：

$$I=(\tau;\phi) \tag{19}$$

其中，τ 和 ϕ 具有如下含义：τ 代表一个二进制向量并能有效地代表一个候选基因子集。每个等位基因 τ_i 代表相应的基因 g_i 被选择 $\tau_i=1$ 或者未被选择 $\tau_i=0$。染色体中的 ϕ 代

表一个实数向量，其中每个ϕ_i与基因g_i的特征向量的判别系数一致。正如第二部分解释的，判别系数代表了基因g_i对基准值w_{opt}的贡献。因此，一条染色体可以表示如下：

$$I=\left(\tau_1,\tau_2,\cdots,\tau_p;\phi_1,\phi_2,\cdots,\phi_p\right) \tag{20}$$

其中τ和ϕ的长度由p和t统一去噪后候选基因数目决定。

需要注意的是，这种染色体编码方式具有更广泛的适用性，且比大多数利用遗传算法进行的特征选择更丰富，就是说，除了候选基因子集之外，染色体还包括其他信息（比如这里的 LDA 判别系数），这些信息在设计强有力的交叉和变异操作时有很大帮助。

③ 适应度评价

在 LDA-GA 方法中基因搜索的目的是寻找“好的”基因子集，它们具有最小的维数，同时具有最高的预测准确性。为了达到这两个目标，设计一个适应度函数，同时考虑到如下指标（虽然有些冲突）。

在评价一条染色体$I=\left(\tau;\phi\right)$时，适应度函数与染色体分类准确度（f_1）和染色体中被选基因数目（f_2）有关。准确地说，f_1是利用 LDA 分类器对训练数据集进行分类，评估基因子集τ的分类准确度得到的，并且正式定义如下：

$$f_1\left(I\right)=\frac{TP+TN}{TP+TN+FP+FN} \tag{21}$$

其中，TP和TN分别代表实际的正常和患病样本，如：准确分类；$FP(FN)$是被误分为正常（患病）样本的患病（正常）的样本数量。

适应度函数的第二部分f_2可以通过下面方程进行计算：

$$f_2(I)=\left(1-\frac{m_\tau}{p}\right) \tag{22}$$

其中，m_τ表示候选基因子集τ中值为“1”的位的个数，如：被选择基因的个数；p表示和去噪后预选基因个数相对应的染色体的长度。

适应度函数f定义为下述的加权求和：

$$f(I)=\alpha f_1(I)+(1-\alpha)f_2(I)\qquad 0\le\alpha\le 1 \tag{23}$$

其中，α是一个参数，其允许我们给f_1和f_2分配一个相对重要因素。给α赋一个大于 0.5 的值，遗传搜索将会向高分类精度的解进行（可能会以选择更多的基因为代价）。相反，

使用小的α值有助于向小的基因子集方向搜索。因此，变化的α会改变遗传算法的搜索方向。

最后，值得注意的是f的取值范围是$[0,1]$；那么，一个f值大的解就比一个f值小的解好。

将临床数据内容利用 LDA-GA 方法筛选后，得到最优生理基因向量，该向量包含各类生理基因中最有代表性的几个基因，能够给予对广义候选分类特征子集的选择的有效贡献。随后，将最优基因生理向量分别加入到第 3 问中去噪后的候选分类特征子集中，得到的每个广义候选分类特征子集，均含有两部分：其一是候选分类特征子集部分，其二是最优基因生理向量。

为了简化问题，本文选取的最优生理基因向量维数是 1，利用 LDA-GA 方法选取的结果为 APC 基因。因此，将 APC 基因分别加入去噪后的每个候选分类特征子集。这样，就可以利用广义候选分类特征子集对 RBF 支持向量机进行训练，通过测试结果的分析和评价，筛选出具有最佳分类效果的广义特征子集，即广义基因“标签”。

5.4.3 模型测试

将附表 2 中所有 35 个候选特征基因子集的维数均增 1，即$D_i' = \{D_i, g_{869}\}, i = 1,2,...,35$，$D_i'$为融入了有助于诊断肿瘤信息的确定基因“标签”的特征基因子集，用他们来分别训练支持向量机。

基于支持向量机对候选分类特征子集进行效果测试，其中“留一法”的测试次数为 40 次，“独立测试实验”的测试次数为 22 次。先将它们的测试效果以误分次数的方式显示。图 17 中 35 个候选特征分类子集分别训练支持向量机，经“留一法”和“独立测试实验”得到其分类效果，如图 17 所示：

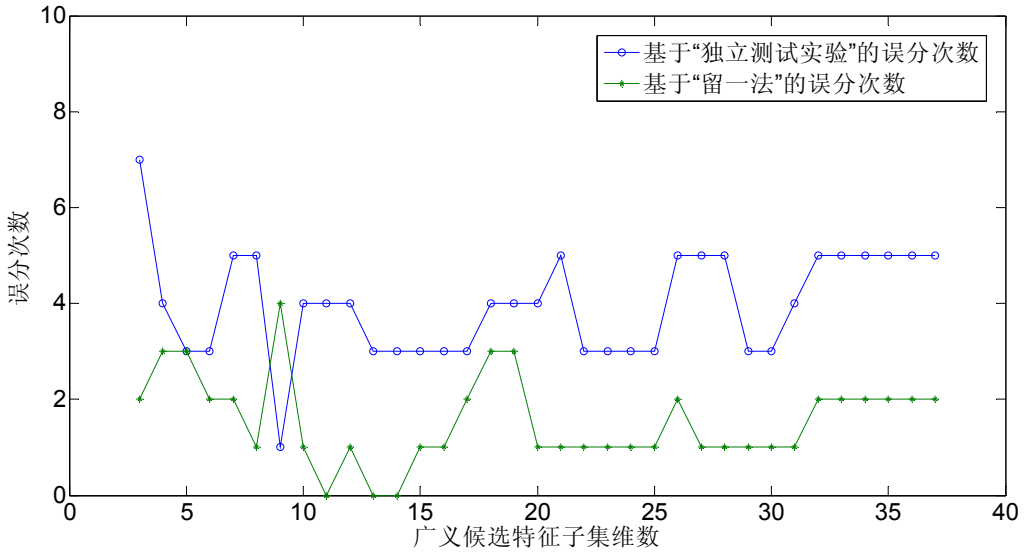

图 17 “留一法”以及“独立测试实验”误分次数

由图 17 并比较图 12 可知，其分类效果与未加有助于诊断肿瘤信息的确定基因“标签”时相比，虽然在总体层面上变化并不明显，但其对低维的候选特征子集的分类效果有提

高作用，下面进一步利用空间对比图展示融入有助于诊断肿瘤信息的确定基因“标签”后的分类效果。

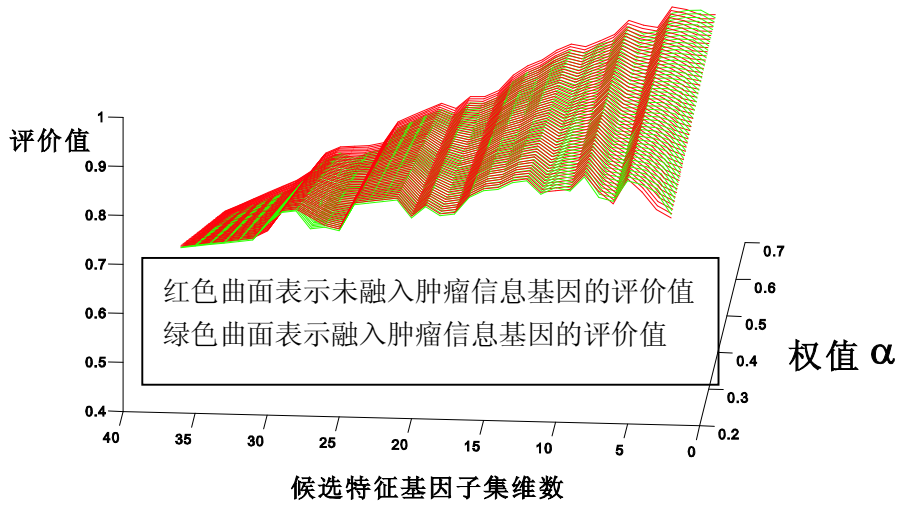

图 18 融入新基因前后基于多指标评价函数的候选特征基因子集评价

由图 18 可知，权值α的变化对于基因“标签”的决定起到了重要的作用，当α较小时，该基因“标签”为 6 维，当α较大时，该基因“标签”为 5 维甚至是 4 维。其评价值始终高于未加有助于诊断肿瘤信息的确定基因“标签”，证明在此种情况下融入有助于诊断肿瘤信息的确定基因“标签”进一步提高了分类的效果。

六、模型评价及改进

1．基于 Bhattacharyya 距离和浮动顺序搜索算法及 SVM 的确定基因“标签”的模型

以 Bhattacharyya 距离作为评价指标，能够将基因表达谱中 2000 个基因缩减至 388 个信息基因，大幅降低了搜索的复杂度，提高了搜索效率，具有较强的剔除无关基因作用。随后，在所得信息基因构成的特征分类子集空间中，利用浮动顺序搜索算法搜索得到维数从 2 到 36 共 35 个候选特征分类子集，利用 SVM 及多指标评价模型筛选出含有 7 个基因的具有较佳分类效果的基因“标签”。利用“留一法”及“独立测试实验”评价并得到最优特征分类子集，其中“留一法”误分次数为 1 次，“独立测试实验”误分次数为 2,维度为 7，最终确定其为基因“标签”，但没有考虑样本的噪声对评价结果的影响。

2．引入 NT_I 及 NT_II 噪声模型进一步提高基因“标签”的分类效果

在进行基因筛选之前先对样本集进行去除噪声的处理，通过分别去除 NT_I 以及 NT_II 两类噪声进一步提高了样本的合理性，之后正常筛选基因。其最终确定的基因标签对样本的“留一法”误分次数为 2，“独立测试实验”误分次数为 3，维数为 5，由于维数得到了降低，因而根据多指标评价模型，其综合评价效果进一步得到了提升。

3．确定广义基因“标签”模型进一步提高基因“标签”的分类效果

在引入噪声模型的基础之上进一步融入这些有助于诊断肿瘤信息的确定基因“标签”，之后正常筛选基因。其最终确定的基因标签对样本的“留一法”误分次数为 3，“独立测试实验”误分次数为 3，维数可以达到 4，由于维数可以进一步降低，因而根据多指标评价模型，其综合评价效果进一步得到的提高。

模型的改进，本模型虽然对基因“标签”的分类效果进行了很大限度的提升，但是仍然有以下内容可以期望在今后进一步的研究中得到完善：1 建立的噪声模型没有考虑到所有可能出现的噪声，需要进一步完善噪声模型。2 当待选择的特征分类子集数量较大时，浮动搜索算法的执行时间会变得很长，如何设计快速的搜索算法是下一步改进的重要方面。

附表 1 不同维数下候选基因特征子集所含基因

维数	候选基因特征子集所含基因
2	22,1772
3	22,1771,1897
4	22,1771,1897,1346
5	22,1771,1897,1346,1750
6	22,698,1897,1346,1924,1740
7	22,698,1897,1346,1924,1740,1352
8	22,698,1897,1346,1924,1740,1352,1423
9	22,698,1897,1346,1924,1740,1352,1423,1522
10	22,698,1897,1346,1924,1740,1352,1423,1522,95
11	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516
12	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252
13	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332
14	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617
15	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531
16	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960
17	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67
18	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186
19	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514
20	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698
21	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453
22	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997
23	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293
24	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853

25	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442
26	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452
27	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849
28	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487
29	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487,1909
30	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487,1909,74
31	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487,1909,74,625
32	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487,1909,74,625,1119
33	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487,1909,74,625,1119,529
34	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487,1909,74,625,1119,529,1154
35	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487,1909,74,625,1119,529,1154,1622
36	1260,377,1897,1346,1924,1740,1352,1423,1522,95,516,252,1332,1617,1531,1960,67,1186,1514,698,453,997,1293,1853,1442,1452,1849,1487,1909,74,625,1119,529,1154,1622,1487

附表 2 加入去噪声处理后不同维数下候选基因特征子集所含基因

维数	候选分类特征子集所含基因序号
2	1582,493
3	1582,493,1346
4	1582,493,1346,586
5	1582,493,1346,586,1870
6	1582,493,1346,1667,1622,1909
7	1582,493,1346,1667,1622,1909,1153
8	549,493,1346,1667,1622,1909,1094,1740
9	549,493,1346,1667,1622,1909,1094,1740,1582
10	1924,493,1346,1667,1383,1909,1073,1740,1582,1917
11	1924,493,1346,1667,1383,1909,1073,1740,1582,1917,1094

12	1924,493,1346,1667,1383,1909,1073,1740,1582,1917,1094,1622
13	1924,493,1346,1667,1383,1909,1073,1740,1582,1917,1094,1622,1439
14	1924,493,1346,1667,1383,1909,1073,1740,1582,1917,1094,1622,1439,686
15	1924,493,1346,1667,1383,1909,1073,1740,1582,1917,1094,1622,1439,686,264
16	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514
17	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889
18	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1918
19	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1917,793
20	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1917,1383,1560
21	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,1853
22	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,1853,1439
23	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,1853,1439,1152
24	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,1853,1439,1152,652
25	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953
26	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560
27	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354
28	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637
29	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637,1717
30	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637,1717,639
31	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637,1717,639,897
32	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637,1717,639,897,682
33	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637,1717,639,897,682,1798
34	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637,1717,639,897,682,1798,1154
35	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637,1717,639,897,682,1798,1154,1635
36	1924,493,1346,1667,1668,1548,1073,1740,1582,1982,1094,1846,997,686,529,1514,1889,1708,1383,1139,549,1439,1152,1378,953,1560,354,1637,1717,639,897,682,1798,1154,1635,830

附程序源代码

1 浮动搜索算法

```
function index_matrix=FFSMLOOP
g1=N1;
g2=N2;
index=zeros(1,50);
index(1,1)=g1;
index(1,2)=g2;
index_matrix=zeros(49,50);
index_matrix(1,1:50)=index(1,1:50);
% t=Bhatt(2,index);
flag_loop=0;
while i<36
    if flag_loop==0
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%
    max_i_1=0;
    max_j=0;
    index_temp1(1,1:50)=index_matrix(i-1,1:50);
    for j=1:388
        flag1=0;
        for k=1:i
            if j==index_temp1(1,k)
                flag1=1;
                break;
            end
        end
        if flag1==0
            index_temp1(1,i+1)=j;
            temp1=Bhatt(i+1,index_temp1);
            if max_i_1<temp1
                max_i_1=temp1;
                max_j=j;
            end
        end
    end
    index_temp1(1,i+1)=max_j;
    index_matrix(i,1:50)=index_temp1(1,1:50);
    end
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%
    index_new=zeros(1,50);
    max1=0;
    max_except_j=0;
```

```
        index_new(1,1:50)=index_matrix(i,1:50);
        for j=1:i+1
            % 第 j 个元素和最后一个进行交换
            index_temp2(1,1:50)=index_matrix(i,1:50);
            temp2=0;
            temp2=index_temp2(1,j);
            index_temp2(1,j)=index_temp2(1,i+1);
            index_temp2(1,i+1)=temp2;
            % index_new(1,i)=0;
            temp3=Bhatt(i,index_temp2);
            if max1<temp3
                max1=temp3;
                max_except_j=j;
            end
        end
        temp4=0;
        temp4=index_new(1,max_except_j);
        index_new(1,max_except_j)=index_new(1,i+1);
        index_new(1,i+1)=temp4; %index_new 即为所求 F'
        index_new(1,i+1)=0;
        %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        flag_loop=0;
        if max1<=Bhatt(i,index_matrix(i-1,1:50))
            i=i+1;
        else
            index_matrix(i-1,1:50)=index_new(1,1:50);
%               index(1,1:50)=index_new(1,1:50);
            if i>2
                flag_loop=1;
                i=i-1;
            end
        end
    end
```

其中 Bhattacharyya 评价函数代码为：

```
function res=Bhatt(n,index) % 计算评价函数 n,为 F 向量的维数，inden 为 50 维向量，里面存的信息为 array_filter1 矩阵的被选中的 n 个行的分别的行数，如果 n 小于 50，则后面的元素为 0
load matlab_data_after_filter02;
u1=zeros(n,1);
u2=zeros(n,1);
mid1=zeros(22,n);
mid2=zeros(40,n);
o1=zeros(n,n);
o2=zeros(n,n);
```

```
for i=1:n
    u1(i,1)=sum(array_filter1(index(i),1:22))/22;
    u2(i,1)=sum(array_filter1(index(i),23:62))/40;
    mid1(1:22,i)=(array_filter1(index(i),1:22))';
    mid2(23:62,i)=(array_filter1(index(i),23:62))';
end
o1=cov(mid1);
o2=cov(mid2);
res=0.125*(u2-u1)'*inv((o1+o2)/2)*(u2-u1)+0.5*log(det((o1+o2)/2)/sqrt(det(o1)*det(o2)));
```

2　去除 NT_I 类噪声算法代码

```
function V_RES=NT_I
load matlab3;
V=zeros(2000,1);
for i=1:2000
    V(i,1)=max(DATA(i,:))/min(DATA(i,:));
end
V_RES=sort(V);
t=1:2000
plot(t,V_RES);
```

3 去除 NT_II 类噪声算法代码

```
function [series_numII,series_array_filterII]=NT_II(del1_num,boundryII)
load matlab_filter;
numII=2000-del1_num;
arrayII=zeros(numII,63);
arrayII(1:numII,1:63)=array_filter011(1:numII,1:63);
times=80;
series_numII=zeros(1,times);
series_array_filterII=zeros(times,numII,63);
nb=7; % 取该值时 注意“%%%%%%%%%%%%%%%%%%%%$$$$$$$$$$$$$$$$$$$$$”处
count1=0;
for i=1:times
    k=1;
    array_temp1=zeros(numII,63);

    count1=count1+1;
    if count1>=20
        count1=0;
        nb=nb-1;
    end
    for j=1:numII
        aa_temp=mod(20+nb*(i-1),62);
        if aa_temp<0.01&&aa_temp>-0.01
```

```
            aa_temp=62;
        end
        %%%%%%%%%%%%%%%%%$$$$$$$$$$$$$$$$$$$$
        if 1+mod(nb*(i-1),62)<aa_temp

temp_V=max(arrayII(j,1+mod(nb*(i-1),62):aa_temp))/min(arrayII(j,1+mod(nb*(i-1),62):aa_temp));
        else

temp_V=max(max(arrayII(j,1:aa_temp)),max(arrayII(j,1+mod(nb*(i-1),62):62)))/min(min(arrayII(j,1:aa_te
mp)),min(arrayII(j,1+mod(nb*(i-1),62):62)));
        end

        if temp_V>boundryII
            array_temp1(k,1:63)=arrayII(j,1:63);
            k=k+1;
        end
    end
    arrayII(1:(k-1),1:63)=array_temp1(1:(k-1),1:63);
    numII=k-1;
    series_numII(1,i)=numII;
    series_array_filterII(i,1:numII,1:63)=arrayII(1:numII,1:63);
end
array_filter_res=zeros(numII,63);
array_filter_res(1:k,1:63)=arrayII(1:k,1:63);
t=1:times;
plot(t,series_numII);
```

4 基于支持向量机及“留一法”测试特征分类子集算法代码

```
function error_times=SVM_C_LOOCV
load matlab_data_after_filter02;
y=zeros(40,1);
for i1=1:14
    y(i1,1)=-1;
end
for i2=15:40
    y(i2,1)=1;
end
error_times=zeros(35,1);
for i=2:36
    x=zeros(40,i+1);
    for j=1:i+1
        x(1:14,j)=(DATA(result_add(i-1,j),1:14))';
```

```
            x(15:40,j)=(DATA(result_add(i-1,j),23:48))';
        end

        for k=1:40
            x1=zeros(40,i+1);
            x1(1:40,1:i+1)=x(1:40,1:i+1);
            y1=zeros(40,1);
            y1(1:40,1)=y(1:40,1);
            temp_x=zeros(1,i+1);
            temp_x(1,1:i+1)=x1(k,1:i+1);
            x1(k,1:i+1)=x1(40,1:i+1);
            x1(40,1:i+1)=temp_x(1,1:i+1);
            temp_y=y1(k,1);
            y1(k,1)=y1(40,1);
            y1(40,1)=temp_y;
            x_train=zeros(39,i+1);
            y_train=zeros(39,1);
            x_train(1:39,1:i+1)=x1(1:39,1:i+1);
            y_train(1:39,1)=y1(1:39,1);
            %%%%%%%%%%%% 代入 libsvm 函数 svm_train 训练支持向量机分类器%%%%%%%%%%%%%
            cmd=['-s 0'];
            model=svmtrain(y_train,x_train,cmd);
            mse_train=zeros(3,1);
            [y_train,mse_train]=svmpredict(y1(40,1),x1(40,:),model);
            error_times(i-1,1)=error_times(i-1,1)+0.5*abs(y1(40,1)-y_train);
        end
    end
```

5 基于支持向量机及“独立测试实验”测试特征分类子集算法代码

```
    function [error_times_test]=SVM_C
    load matlab_data_after_filter02;
    y=zeros(40,1);
    yt=zeros(22,1);
    for i1=1:14
        y(i1,1)=-1;
    end
    for i2=15:40
        y(i2,1)=1;
    end
    for i3=1:8
        yt(i3,1)=-1;
    end
    for i4=9:22
```

```
        yt(i4,1)=1;
    end
    test_result=zeros(22,35);
    test_mse=zeros(3,35);
    error_times_test=zeros(35,1);
    for i=2:36
        x=zeros(40,i+1);
        for j=1:i+1
            x(1:14,j)=(DATA(result_add(i-1,j),1:14))';
            x(15:40,j)=(DATA(result_add(i-1,j),23:48))';
        end
        %%%%%%%%%%%% 代 入 libsvm 函 数 svm_train 训 练 支 持 向 量 机 分 类
器%%%%%%%%%%%%%%%
        cmd=['-s 0'];
        % cmd=['-c ',num2str(C),' -g ',num2str(s),' -s 3 ' ,'-p ',num2str(e)];
        model=svmtrain(y,x,cmd);
        %%%%%%%%%%测试集检验%%%%%%%%%%%%%%%%%%%%%
        xt=zeros(22,i+1);
        for k=1:i+1
            xt(1:8,k)=(DATA(result_add(i-1,k),15:22))';
            xt(9:22,k)=(DATA(result_add(i-1,k),49:62))';
        end
        [test_result(:,i-1),test_mse(:,i-1)]=svmpredict(yt,xt,model);
        for m=1:22
            error_times_test(i-1,1)=error_times_test(i-1,1)+0.5*abs(test_result(m,i-1)-yt(m,1));
        end
    %       test_mse(1,i-1)=a;
    end
```

参考文献

[1] 李颖新，刘全金，阮晓钢. 急性白血病的基因表达谱分析与亚型分类特征的鉴别，中国生物医学工程学报，Vol. 24 No. 2 :240-244，2005

[2] Duda OR，Hart PE，Stork GD. Pattern Classification [M] . Second Edition. New York :John wiley & Sons 2001 :46-48.

[3] Theodoridis S，Koutroumbas K. Patter Recognition [M]. Second Edition. New York :Academic Press，2003，177-179.

[4] Padil P，Novovicova J，Kittler J . Floating search method in feature selection [J]. Pattern Recognition Letters，1994，15 (11) : 1119-1125.

[5] 刘全金，李颖新，阮晓钢. 基于基因表达谱的结肠癌特征基因选取，昆明理工大学学报，Vol.31 No.1:89-92.

[6] 包雷，黄英武，孙之荣. 基于基因表达谱的肿瘤分型和特征基因选取，生物物理学报， Vol.18 No.4:413-417，2002

[7] 宣国荣，柴佩琪. 基于 Chernoff 上界的特征选择，模式识别与人工智能，Vol.9 No.1:26-30，1996
[8] Vapnik V. Statistical Learning Theory[M]. New York: Wildy, 1998.
[9] 张浩然，韩正之，李昌刚. 支持向量机，计算机科学，Vol.29 No.12:135-138，2002
[10] Z. Sun, P. Yang, Gene expression profiling on lung cancer Outcome Prediction: Present Clinical Value and Future Premise, Cancer Epidemiology Biomarkers & Prevention, 2006, 15(11): 2063-2068
[11] 张毅，姜军. 抑癌基因失活在多原发性恶性肿瘤发生中的意义，中国普外基础与临床杂志，Vol.7 No.6:418-420，2000
[12] 盖文君，姚宏. 子宫平滑肌肿瘤 P16 抑癌基因表达的定量研究，山西医药杂志，Vol.36 No.7:597-598，2007
[13] 唐卫中，高枫，唐宗江. 大肠良恶性肿瘤 PTEN/MMAC1/TEP1 肿瘤抑制基因的蛋白表达研究，Vol.7 No.2：9-13，2001