全国第七届研究生数学建模竞赛

题 目 确定肿瘤的重要基因信息

摘 要：

随着生物分子学的发展，人们已经发现癌症与基因之间存在密切关系。

本文通过对比基因表达谱中不同样本的表达水平的差别，区分出无关基因和信息基因，在信息基因的空间中搜索分类能力强的特征子集,然后将所有样本划分为训练集和测试集，使用支持向量机 SVM 和人工神经网络检查特征子集的分类能力。其次，再考虑噪声的定义，以及如何去除噪音，并分析噪声对特征子集分类能力的影响。最后，生成特征子集的时候需要考虑已有的医学发现，提出基于知识库的基因图谱分析模型 KFS 模型，有效利用了医学成果。

第一问，利用巴氏距离模型和理想基因模型区分无关基因和相关基因，剔除基因表达谱中无关基因，达到降维的效果。

第二问，使用 FSSM 算法在信息基因空间里寻找出候选特征子集，然后将样本划分为训练集和测试集，使用支持向量机 SVM 和人工神经网络，对 FSSM 搜索出来的特征子集的分类能力进行判定。本文得到由 5 个信息基因组成的特征子集，样本分类正确率达到 95.79%。

第三问，考虑了阈值滤波和主成分分析两种去噪模型，并阐述噪音模型在高斯过程分类器的构建中的作用，最后论述噪音能够在学习算法中防止过渡拟合从而可以孵化出泛化能力更强的分类器用于确定基因标签。

第四问，提出基于知识库的基因图谱分析模型 KFS 模型，该模型在引入信息基因知识库的基础上，对基因图谱进行去噪处理、样本评价函数增益、剔除无关基因，并采用基于知识库的 KFSSM 算法获得特征子集，最后分别采用 SVM 及 ANN 方法获取信息基因集合。本文最后对已知临床经验的结肠癌数据进行处理得到一组信息基因组合，样本分类正确率达到 94.52%。

关键词：基因表达谱，信息基因，巴氏距离，FSSM，噪声，KFS 模型

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，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。

面对提取基因表达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表达谱信息就行。

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。

二、基本的模型假设

1、基因表达谱中的样本类别没有错误。
2、基因表达谱中的癌症病人样本都是结肠癌病人。
3、基因表达谱中有重复的基因标签，我们假定重复的基因标签的样本分类能力类似，所以只处理其中一个基因标签
4、所有的癌症病人都处于同一时期，不分早期和晚期。
5、基因表达谱中有一些重复的基因标签，比如 HSAC07、UMGAP 和 i 都出现了 4 次，而 Has.13491、Has.44472 等基因出现了两次。我们假定每一个标签都具有代表性，所以我们只处理一个基因标签，而不考虑其他重复的基因标签。这样的话，基因表达谱信息表中总共有 1911 个不同类型的基因标签。本文剩下所有的数据处理都是针对这 1911 个基因标签。

三、名词解释

基因表达谱：关于基因表达的信息，可以看成是一个矩阵或者一个向量，矩阵或者向量元素的数值大小就是该基因的表达水平。
基因表达水平：可以理解为样本中某种基因的数量或者密度。
理想基因：一种可以完全分辨出样本类别的基因，它在不同类型的样本中的表达水平相差很大。
信息基因：可以决定样本类别的一组基因。
无关基因：有一些基因的表达水平在所有的样本中非常接近，对样本分类没有帮助的基因。

特征子集：每一种信息基因的组合。
信噪比：作为衡量基因对样本分类贡献大小的量度。
训练集：用来训练分类器学习能力的样本集，包含正常人的样本和结肠癌病人的样本。
测试集：用来测试分类信息基因分类能力的样本集，包含正常人的样本和结肠癌病人的样本。
分类器：通过学习训练集中样本之后，可以自动的对给定的测试集中样本进行分类的一类程序。

四、符号化

Bi：基因标签 i 的巴氏距离（1≤i≤1911）。
S_N：无关基因集合。
S_I ：信息基因集合。
e：理想基因，与肿瘤有很强的关联性。
Num(S_N)：无关基因集合的大小。
Num(S_I)：信息基因集合的大小。
正常样本 nSi：第 i 个正常人样本（1≤i≤22）。
病人样本 cSj：第 j 个结肠癌病人样本(1≤j≤40)。
基因表达谱信息：A[m×n]，用一个 m×n 的矩阵来表示基因，其中 m=62，n=1911
vij：样本 i 在基因 j 上的表达水平。
基因矢量 Vg：基因 g 在各个样本上的表达水平的一个矢量，Vg={vg1, vg2,…, vgn}。

五、模型的建立与求解

5.1 问题一的分析、建模与求解

5.1.1 问题一的分析

目前人们通过生物芯片的技术可以快速检测样本的基因表达水平，人类各类组织的正常基因表达谱已经获得，但是还没有有效的方法能够定位与癌症直接有关的基因。

问题一的出发点在于，人类基因表达谱中包含有太多与癌症无关的基因，这大大地增加了人们从基因表达谱中搜索与癌症相关的信息基因的难度，而一般情况下，人们认为直接与特定类型癌症相关的突变基因数目很少，所以我们应该考虑首先从基因图谱中删除大量的无关基因，缩小搜索信息基因的范围。这个步骤可以称为基因表达谱去除无关信息的降维处理。本题我们从下面两个角度对基因表达谱初步降维：

1、信息基因在不同类型样本表达水平的差异。

2、基因与理想基因的相似度。

我们首先从癌症的“基本致病机理”角度分析肿瘤与基因之间的可能存在关系。癌症会导致信息基因在不同类型样本的表达水平上产生一些差异。我们应当用一种合理的指标将这种差异量化。目前比较通用的方法是比较样本的平均值和方差，我们考虑一种模型，可以综合考虑这两方面的因素。

其次，我们假设有一种理想基因，这种基因在不同类型上的样本上的表达水

平差异非常大。我们通过样本在理想基因上的表达水平就可以直接判断样本的类型。利用这种理想基因，我们通过比较基因表达谱中各个基因与理想基因的相似度。相似度高的基因可以认为是信息基因，相反，如果某个基因标签与理想基因的相似度很小，我们基本可以认为它是无关基因。

本文在处理第一题时，综合利用了巴氏距离模型和理想基因模型，以巴氏距离模型为主，但是由于基因表达谱中的噪声会影响巴氏距离模型选出来的信息基因的效果，所以再以理想基因模型为辅，选取一些与理想基因相似度高的基因，防止将一些信息基因剔除。最后选取大小为 250 的信息基因集合，大约占题目所给基因总数的 20%，作为第二问的特征子集的搜索空间。

下面详细介绍我们建立的降维模型。

5.1.2 问题一的模型建立

我们首先考虑下如何量化信息基因在不同类型样本中表达水平的差异以及如何利用这种差异将基因分类，区分出无关基因和信息基因。

Golub 等人以“信噪比”(Signal to noise ratio)[1]指标作为衡量基因对样本分类贡献大小的度量，信噪比的定义如下：

$$d = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \qquad (1)$$

其中：d 是基因的信噪比，μ 1 和 μ 2 分别是该基因在两种样本中表达水平的均值，σ 1 和 σ 2 是该基因在两种样本中表达水平的标准差。

但是使用这种方法来区分信息基因和无关基因存在应的问题。如果 d=0，该基因就会被当做无用基因删除，而实际上，如果该基因在两种样本中表达水平方差有很大差异，那么很有可能这个基因与癌症有很密切的关系。所以，我们需要选用一个模型，综合考虑平均值和方差的差异。

5.1.2.1 巴氏距离模型

巴氏距离既考虑到基因在样本中均值，也考虑到基因在样本中的方差分布，是一个很好的信息度量指标。它综合考虑了均值和方差差异对样本分类的作用。它的定义如下：

$$B = \frac{1}{4}\frac{(u_1 - u_2)^2}{(\sigma_1^{\ 2} + \sigma_2^{\ 2})} + \frac{1}{2}\ln(\frac{\sigma_1^{\ 2} + \sigma_2^{\ 2}}{2\sigma_1\sigma_2}) \qquad (2)$$

其中 B 为基因的巴氏距离。由式(2) 知，巴氏距离由两部分构成：第一项体现了基因在两个类别中分布均值的差异对样本分类的贡献；第二项体现了分布方差的不同对分类的贡献。依据该距离公式,即使基因在两类不同样本中分布的均值相同，只要分布的方差出现大的差异，仍然可以获得较大的距离值 [3]。

从模式分类的角度来看，基因的巴氏距离越大，说明该基因的分类能力越强，基因的分类信息越多。设 S_N 是无关基因集合，S_I 是信息基因集合，我们设置一个阈值 θ，巴氏距离大于 θ 可以认为是无关基因，巴氏距离小于 θ 的可以认为是信息基因。

$$g \in \begin{cases} S_I B(g) > \theta \\ S_N B(g) \le \theta \end{cases} \qquad (3)$$

其中，g 是基因，B(g)为基因 g 的巴氏距离，θ 是选取的巴氏距离的阈值。

利用公式 3，选取好的阈值，我们就可以区分出无关基因集合 S_N 以及信息基因集合 S_I。

5.1.2.2 理想基因模型

当然我们不能只从巴氏距离这一个标准来衡量基因分类信息的大小，同时由于生物基因芯片不可避免地存在一些噪声（噪声的处理方法会在第三问中提出解决方案），噪声会对样本的均值和方差产生较大的影响，而均值和方差是计算基因巴氏距离的两个重要因素。所以为了使基因的分类更为合理，除了以巴氏距离，我们还提出理想基因的概念。

题目中给定的基因表达谱数据可以用矩阵 A[m×n]表示，其中 m 表示基因的数目，去除重复的基因之后，只有 1911 个，n 表示样本的数量，总共有 62 个。矩阵中元素 vij 表示第 j 个样本对基因 i 的表达水平。我们首先对基因表达谱的数据作归一化处理，使得矩阵 A 中每个元素的值都在[-1, 1]之间：

$$v_{ij}^{'}=\frac{2v_{ij}-v_{\max}-v_{\min}}{v_{\max}-v_{\min}},0\le i\le 1911,0\le j\le 62 \qquad (4)$$

其中 v_{max} 是指矩阵 A 中的最大值，v_{min} 是指矩阵 A 中的最小值，$v_{ij}^{'}$ 是 v_{ij} 归一化之后的数值。

本题中的样本总数有 62 个，分为两类，一类是正常人样本，另一类是结肠癌患者样本。正常人样本总共有 22 个，样本编号从 N1—N2，结肠癌患者样本有 40 个，样本编号从 C1—C40。基因 g 在每个样本中表达水平 Vg={vg1, vg2,…, vgn}可以看成是一维向量。我们将理想基因 e 定义为：

$$v_{e1}=v_{e2}=...=v_{e22}=-1,v_{e23}=v_{e24}=...=v_{e62}=1 \qquad (5)$$

如果基因 g 是结肠癌的信息基因，它本身携带有分类信息越多，分类能力越强，那么它越接近于理想基因 e。我们从两个方面来考虑基因 g 向量与理想基因 e 的接近程度，一个是基因向量之间的夹角的余弦值，如果余弦值靠近-1 或者 1，那么这两个基因向量夹角越小。此外，这两个向量之间的欧拉距离也可以作为一个衡量的标准，两个基因向量之间的欧拉距离越小，说明这两个基因向量越靠近。最后，我们用相似度这个值来量化信息基因与理想基因之间的接近程度[8]。

（1）基因 g 与基因 e 的夹角定义为 θ_{ge}:

$$\cos\theta_{ge}=\frac{V_g\bullet V_e}{|V_g||V_e|} \qquad (6)$$

其中 $|V_g|=\sqrt{\sum_{i=1}^{m}v_{gi}^{\ 2}},|V_e|=\sum_{i=1}^{m}v_{ei}^{\ 2}$,

（2）基因 g 与基因 e 的欧拉距离为：

$$D_{ge}=\begin{cases}\sqrt{\sum_{i=1}^{m}(v_{gi}-v_{ei})^2}, cos\theta_{ge}>0\\ \sqrt{\sum_{i=1}^{m}(v_{gi}+v_{ei})^2}, \cos\theta_{ge}\le 0\end{cases} \quad (7)$$

（3）基因 g 与基因 e 的相似度为：

$$S_{ge}=\frac{|\cos\theta|}{D_{ge}} \quad (8)$$

从相似度的定义中，我们可以看出，基因 g 与理想基因 e 向量的欧拉距离不变，夹角越小，余弦值的绝对值越大，相似度越大。基因 g 和 e 的夹角不变，欧拉距离越小，相似度越大。相似度定义很好地量化了基因 g 与理想基因 e 之间的相关性，相似度越高，说明基因 g 的分类能力越强。同样的，我们可以通过设置适当的阈值 θ，将基因表达谱中信息基因和无关基因区分开，达到降维的效果。

$$g\in\begin{cases}S_I, S_{ge}>\theta\\ S_N, S_{ge}\le\theta\end{cases} \quad (9)$$

5.1.2.3 综合模型

本文在处理第一题时，综合利用了巴氏距离模型和理想基因模型，以巴氏距离模型为主，但是由于基因表达谱中的噪声会影响巴氏距离模型选出来的信息基因的效果，所以再以理想基因模型为辅，选取一些与理想基因相似度高的基因，防止将一些信息基因剔除。

我们首先计算基因的巴氏距离前 200 的基因，然后再计算与理想基因的相似度，选取相似度值前 50，并且不与前面重复的基因，组成大小为 250 的信息基因集合。这个基因集合大约占题目所给基因总数的 20%，大幅压缩了冗余基因。这 250 个基因集合作为第二问的特征子集的搜索空间。

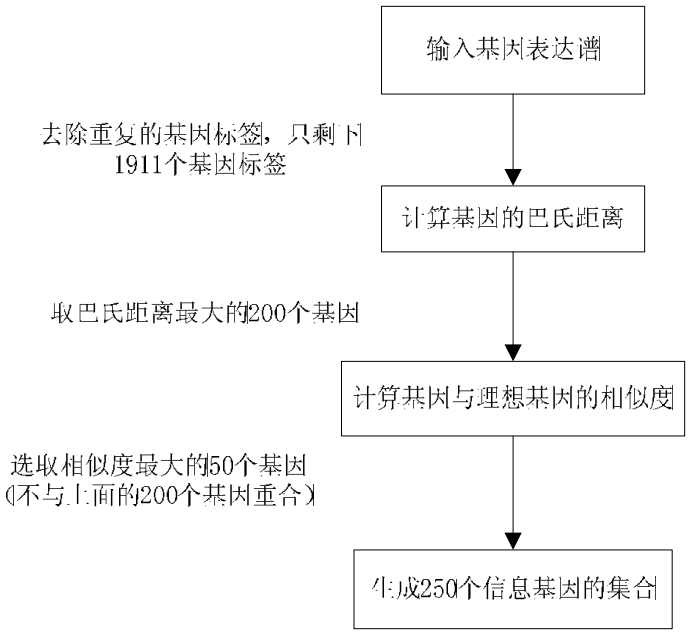

图 5.1.1 综合模型的分类基因流程图

下图是综合模型得到的信息基因集合的构成。

基因表达谱

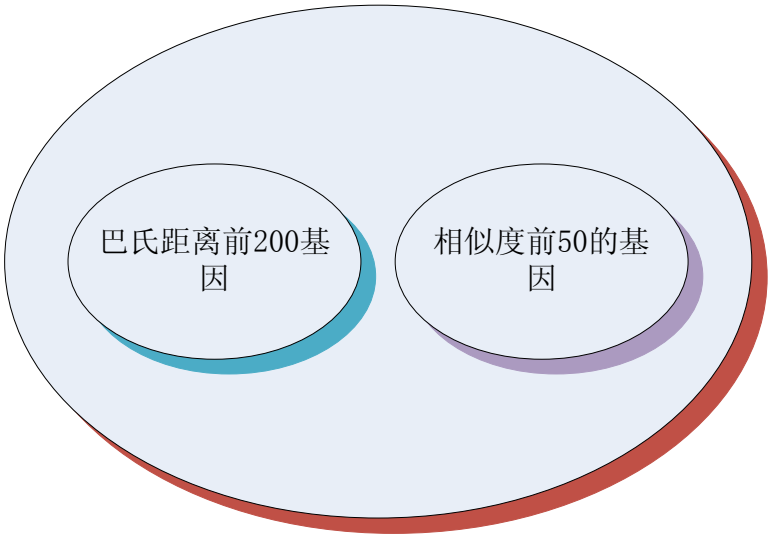

图 5.1.2 综合模型的分类基因组成

5.1.3 问题一的模型求解

1、计算所有基因标签的巴氏距离。

根据巴氏距离计算公式，我们得到了 1911 个基因的巴氏距离分布情况，见表 2。

表 5.1.2 基因标签的巴氏距离分布

巴氏距离	基因个数	百分比
0~0.05	1492	78.04%
0.05~0.1	290	15.18%
0.1~0.15	78	4.08%
0.15~0.2	31	1.62%
0.2~0.25	13	0.68%
0.25~0.4	7	0.36%

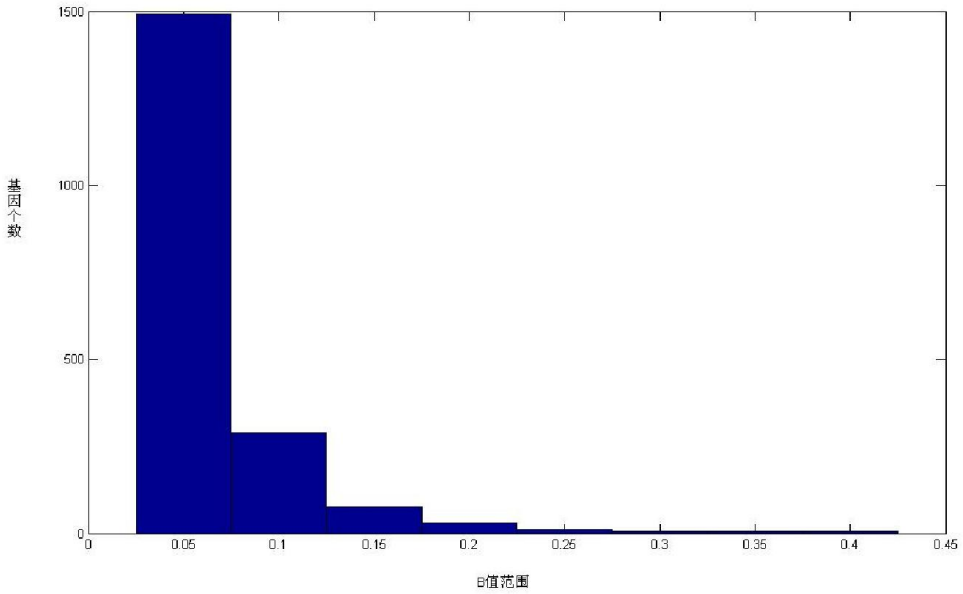

图 5.1.3 基因标签的巴氏距离分布直方图

2、计算剩余基因与理想基因的相似度

接下来，我们基因表达谱中所有基因与理想基因的相似度，然后取相似度前50，且不与巴氏模型的结果重复的基因。

表 5.1.3 基因的相似度分布表

相似度	基因数目	百分比
0.0-0.02	638	33.38%
0.02-0.04	934	48.87%
0.04-0.06	291	15.23%
0.06-0.08	43	2.25%
0.08-0.10	5	0.26%

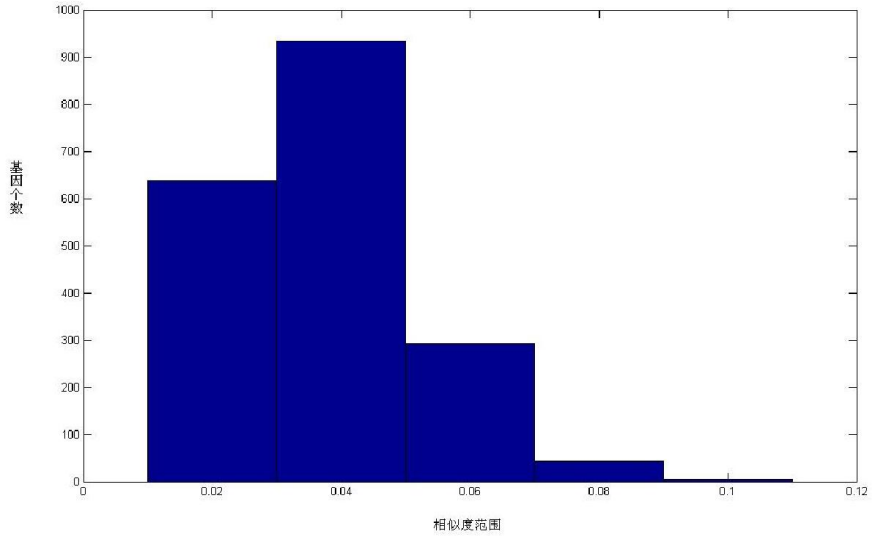

图 5.1.4 基因标签的相似度分布直方图

5.1.4 问题一的结果及分析

问题一的主要目的在于剔除与肿瘤无关的基因，通过巴氏距离模型和理想基因模型处理之后，基因的数量从原来的 1911 个大幅度地减少到了 250 个，降低了第二问 FSSM 算法的搜索特征子集的空间。

5.2 问题二的分析、建模与求解

5.2.1 问题二的分析

问题一和问题二其实都是对基因降维，问题一从单个基因分类能力的角度出发，剔除癌症无关基因，缩小了人们搜索与癌症相关的信息基因的范围。而问题二则是建立在问题一的基础上，从剩下的 250 个信息基因中搜索特征子集，而这 250 个基因可以组成 2^{250} 个不同的特征子集，这就需要一个很高效的搜索算法，同时也需要有一个合理的评价函数，能够评价不同的特征子集的分类能力，从而筛选出分类能力强的特征子集。最后将样本分成训练集和测试集，再选择合适的具有学习能力的分类器，查看候选特征子集的分类能力。特征子集大小和分类准确率为评价指标可以作为衡量指标。

5.2.2 基于 FSSM 算法的特征子集的生成

本文采用 FSSM(Floating Sequential Search Method)搜索算法[2][3]，对特征子集所构成的子空间进行搜索，从中选取 30 个具有不同维数的待选分类特征子集，然后使用 SVM 和人工神经网络检验这 30 个特征子集的分类能力。

FSSM 搜索算法中采用关键函数 J 作为动态搜索特征子集过程中的评价函数，评价函数的描述如下[3]：

$$J(F_i)=\frac{1}{8}(\mu_2-\mu_1)^T\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_2-\mu_1)+\frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \qquad (10)$$

其中，Fi 表示含有 i 个信息基因的特征子集，μ1 、μ2 表示特征子集 Fi 中的信息基因在正常样本和结肠癌样本中的均值向量，Σ1、Σ2 表示 Fi 中的信息基因在正常样本和结肠癌样本中数据的的协方差矩阵，本文中Σ1 为 i×22 矩阵，Σ2 为 i×40 矩阵。

J 函数也是基于巴氏距离的，不过计算的是信息基因集合的巴氏距离。第一问我们利用巴氏距离模型，计算出单个基因的巴氏距离，很好地区分出信息基因和无关基因。FSSM 中 J 函数的作用在于，它从同样大小的信息集合中选择 J 值较大的，具有最强的分类能力子集代表。

FSSM 算法中，令数组 Fmax[i]表示含有 i 个信息基因的具有最大评价函数值的特征子集，本文采用 matlab 语言实现了 FSSM 算法，数组 SelectMax [i]表示计算过程中，计算出含有 i+1 个基因的最大特征子集时选择 i 个的特征子集的最大 J 值，参考文献[2]中的算法思想给出 matlab 算法的处理过程如图 2.1 所示。

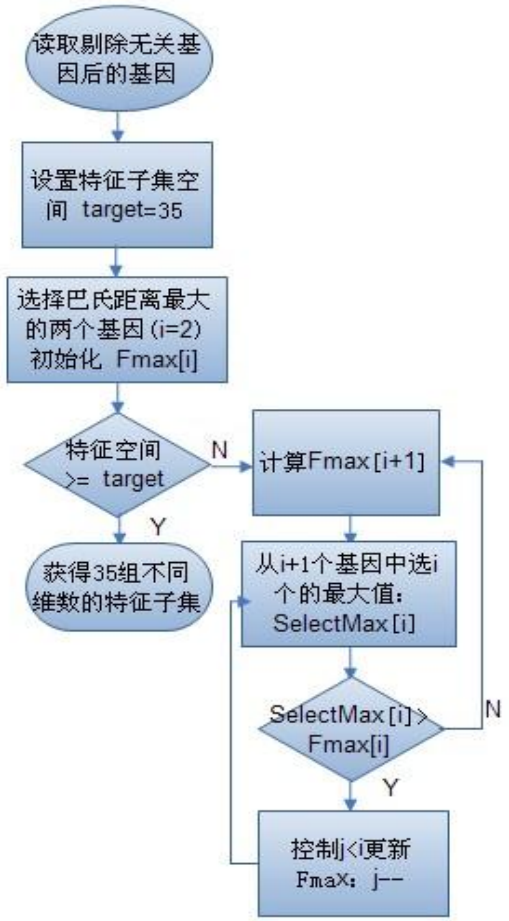

图 5.2.1 FSSM算法matlab实现的算法流程

通过运行FSSM算法，最终生成30个信息基因的特征子集，本文的下一节分别通过支持向量机方法(SVM)和人工神经网络方法考察选出的特征子集的分类能力。

5.2.3.1 基于支持向量机（SVM）的分类特征子集选择

本文的上一节采用FSSM算法生成了34个具有不同维数的特征子集，该部分以支持向量机为分类器对34个特征子集进行样本识别，最终获得具有最大分类正确率的基因组合。

支持向量机是一种基于统计学习理论，采用结构风险最小化原理的机器学习算法。机器学习的目的是根据给定的训练样本对输入输出之间的依赖关系的估计，使得可以对未知的输出尽可能准确的预测。支持向量机的核心思想就是调整评价函数使得最好地利用边界样本点的分类信息，从而构造出最佳分类超平面,因此支持向量机可以获得很好的泛化能力并且与样本的具体分布无关[7]。

结合基因图谱数据特点发现，支持向量机可以有效地处理高维样本的分类问题，计算复杂度受样本维数的影响较小，适合处理小样本、高维数的基因表达谱数据的样本分类问题。

本文使用34组具有不同维数的特征子集依次作为特征属性使用支持向量机学习出分类模型后验证其分类能力。由于样本实例的数目有限，我们采用10-fold交叉验证的方法来评估学习得到的模型的分类能力，进而评估选择的特征基因组合的识别能力。具体方法是将数据集分成10份，轮流将其中9份做训练1份做测试，10次的结果的均值作为对算法精度的估计。

5.2.3.2 基于人工神经网络（ANN）的分类特征子集选择

人工神经网络是一种应用类似于大脑神经突触联接的结构进行信息处理的数学模型. 在这一模型中, 大量的节点之间相互联接构成神经网络，以达到处理信息的目的。人工神经网络对矢量进行样本识别前需要进行训练，训练的过程就是应用一系列输入矢量，通过某种算法逐步调整权值和阈值的过程，通过训练人工神经网络对一组输入矢量产生希望的输出。训练后的人工神经网络即可以用于对正常样本和结肠癌样本的分类。

本文也采用了基于人工神经网络的分类器，使用34组具有不同维数的特征子集依次作为特征属性使用支持向量机学习出分类模型后验证其分类能力，进而评估选择的特征基因组合的识别能力。可以与SVM的结果进行对比。

5.2.4 问题二的结果及分析

本题我们首先使用FSSM方法从250个信息基因中搜索，生成34个不同大小的特征子集，子集大小从2到35不等。然后在62个样本集合上使用SVM（支持向量机）和人工神经网络算法训练出分类器，然后评估分类器的分类能力，进而评估所用基因特征子集识别

下面给出经过SVM100次10-fold交叉验证之后，样本分类准确率在前五的特征子集以及它们的分类准确率。最好的一组特征子集的样本分类能力达到了95.79%，并且该组特征子集的大小只有5。符合我们之前对特征子集的分类能力以及子集大小的要求。

表5.2.1 特征子集分类准确率前五名

特征子集	大小	分类准确率
Hsa.37937, Hsa.710, Hsa.3016, Hsa.5392, Hsa.6080	5	95.79%
Hsa.37937, Hsa.710, Hsa.3016, Hsa.5392	4	94.32%
Hsa.37937, Hsa.549, Hsa.710, Hsa.3016, Hsa.5392, Hsa.6080	6	91.10%
Hsa.37937, Hsa.549, Hsa.710, Hsa.3016, Hsa.5392, Hsa.6080, Hsa.2058, Hsa.43331, Hsa.8214, Hsa.823, Hsa.957, Hsa.33965, Hsa.816, Hsa.490, Hsa.732, Hsa.36689, Hsa.2928, Hsa.8147, Hsa.6814, Hsa.2250	20	89.56%
Hsa.37937, Hsa.549, Hsa.710, Hsa.3016, Hsa.5392, Hsa.6080, Hsa.2058, Hsa.43331, Hsa.8214, Hsa.823, Hsa.957, Hsa.33965, Hsa.816, Hsa.490, Hsa.732, Hsa.36689, Hsa.2928, Hsa.8147, Hsa.6814, Hsa.2250, Hsa.7048, Hsa.58	22	89.27%

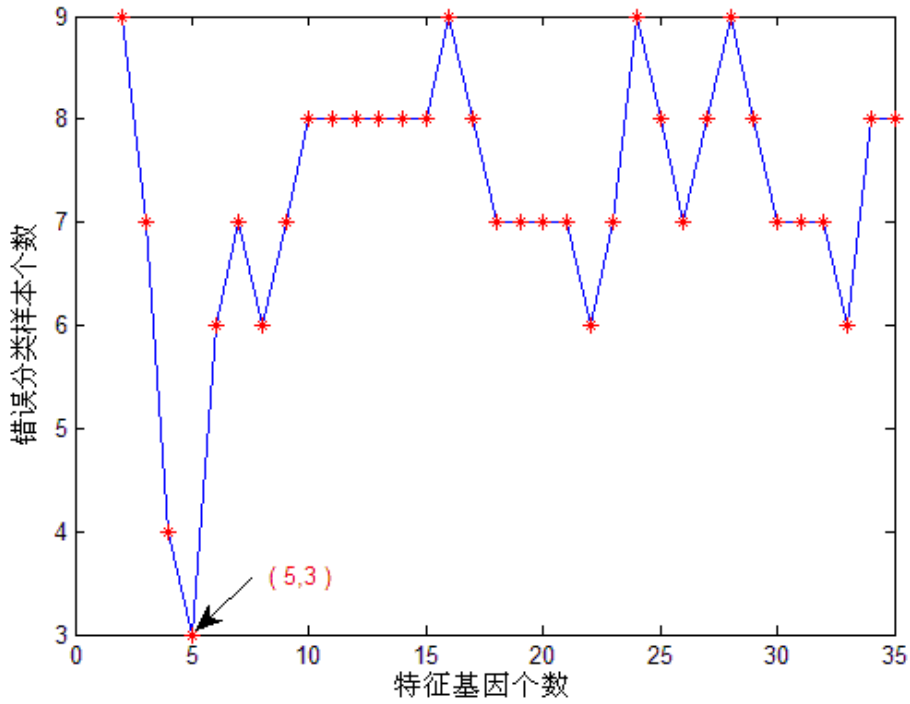

图5.2.2 SVM下不同特征子集的样本识别情况

为了验证最优特征子集的准确性，我们使用Matlab工具绘制了Hsa.37937，Hsa.710，Hsa.3016，Hsa.5392，Hsa.6080这五个基因在62个样本中的表达水平。说明一下，下面的五张图中，*代表正常人样本的基因表达水平，+代表癌症病人样本的基因表达水平。

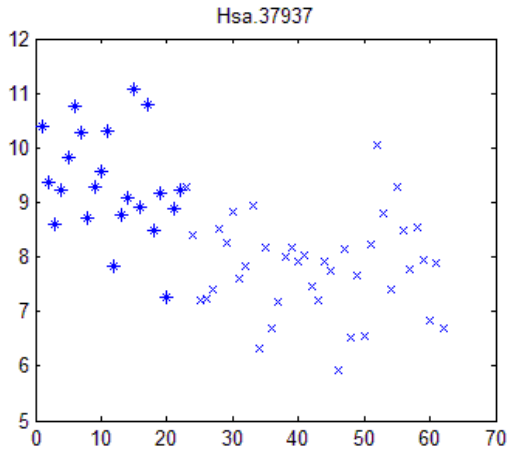

图5.2.3 基因标签Has.37937在不同样本的表达水平

通过这张图，我们可以明显地看出Has.37937基因在正常人和结肠癌病人两类样本的表达水平差异非常明显。

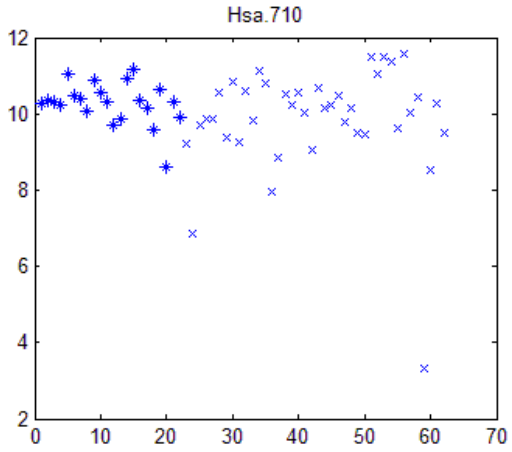

图5.2.4 基因标签Has.710在不同样本的表达水平

这张图显示基因Has.710的分类效果其实并不理想，无论从均值还是从方差来看，都不显著。它被选进特征子集的原因既有可能是噪声的影响，这点我们会在第三问中继续讨论。

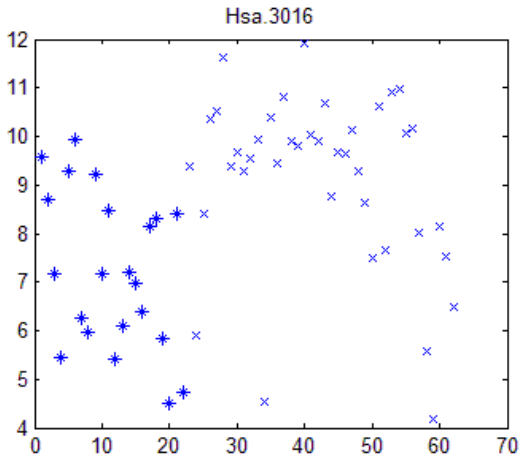

图5.2.5 基因标签Has.3016在不同样本的表达水平

基因标签Has.3016的在两类样本中的表达水平差异也很大。

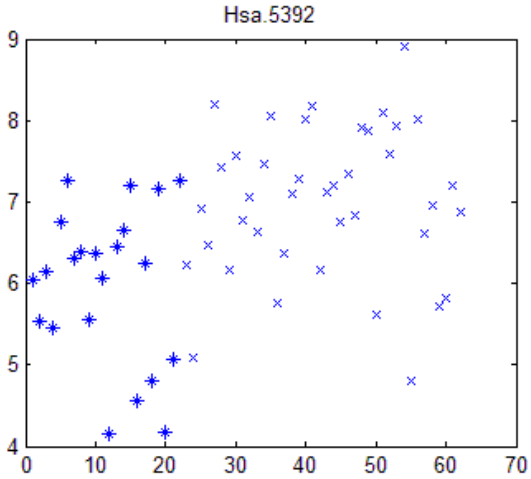

图5.2.6 基因标签Has.5392在不同样本的表达水平

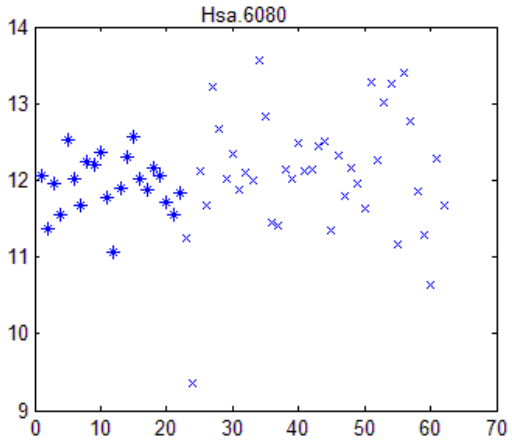

图5.2.7 基因标签Has.5392在不同样本的表达水平

同样的，我们也利用人工神经网络20次循环验证特征子集的分类能力，下表给出了样本分类准确率在前三的特征子集以及它们的分类准确率。通过与前面的SVM的样本分类能力对比，我们发现{Hsa.37937, Hsa.710, Hsa.3016, Hsa.5392, Hsa.6080}这组特征子集的分类能力确实最强，所以确定结肠癌的基因标签就是这组特征子集。

表5.2.2 人工神经网络中特征子集分类准确率前三名

特征子集	大小	分类准确率
Hsa.37937, Hsa.710, Hsa.3016, Hsa.5392, Hsa.6080	5	93.54%
Hsa.37937, Hsa.549, Hsa.710, Hsa.3016, Hsa.5392, Hsa.6080, Hsa.43331, Hsa.8214, Hsa.823	9	93.54%
Hsa.37937, Hsa.549, Hsa.710, Hsa.3016, Hsa.5392, Hsa.6080	6	91.94%

5.3 问题三的分析、建模与求解

5.3.1 问题三的分析

使用 DNA 微阵列（DNA Microarray）技术测量基因表达水平得到的数据具有噪声强、波动大的特点，同时在大量数据的背后还有很多相关变量不能被直接观察到[4]，所以我们有必要仔细研究表达谱中的噪音，并给出相应的理论观点，处理思路和方法。

首先，从提取基因表达谱的角度看，噪声可以分为两类：一种噪声可以认为测量误差引入的噪声，在任何测量过程中无法避免的；另一种噪声是无关基因，该类基因在正常样本和结肠癌样本中的表达水平非常接近，没有为肿瘤的判断提供有用的信息，该类基因的存在增加了提取信息基因的难度。

无关基因是本文研究的问题之一，已经在第一问中试图解决了。下面主要关注第一种噪音。这类噪音主要是测量误差引起的，我们知道测量误差可以分为以下三类[5]：

1. 系统误差，主要是由于测量设备的缺陷、测量环境变化、测量时使用的方法不完善、所依据的理论不严密或采用了某些近似公式等造成的误差。

2. 随机误差，在同一测试条件下，多次重复测量同一量时，误差大小、符号均以不可预定的方式变化着的误差上。

3. 疏失误差，是指在一定的测量条件下，测得的值明显偏离其真值，既不具有确定分布规律，也不具有随机分布规律的误差。

从误差的定义可以知道，系统误差和随机误差是必然存在的。而且在 DNA 微阵列技术中系统误差是比较大的。随机误差是时刻存在的，且其服从一种自然分布。系统误差与随机误差的划分是相对的，二者在一定条件下可以相互转化，即同一误差，既可以是系统误差，又可以成为随机误差。疏失误差是由于测试人员对仪器不了解、或因思想不集中、粗心大意导致错误的读，使测量结果明显地偏离了真值。

对这三种误差的处理方法不同：对于含有疏失误差的测量值应予以剔出；对于随机误差的影响用统计的方法来消除或减弱；对于系统误差则主要靠测量过程中采取一定的技术措施来削弱或对测量值进行必要的修正来减弱其影响。结合到基因图谱信息提取这项具体技术中，我们无法求证系统误差的大小，只能结合到系统误差和随机误差的关系将其中的一部分作为随机误差来处理。而疏忽误差和设备老化、环境突变等恶劣因素引发的系统误差导致的是异常数据，会影响到建模方法的有效性，所以是我们需要在数据中剔除或者修正的。

下面我们针对疏忽误差和随机误差引发的噪声分别建立模型。

5.3.2 问题三的模型建立

我们对两种噪声分别讨论与之对应的模型：

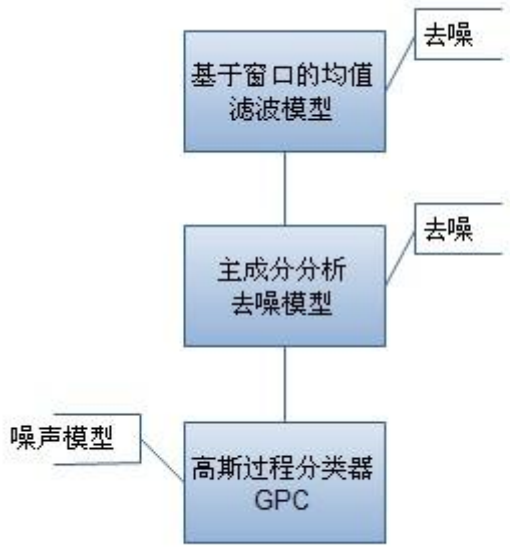

图 5.3.1 去噪模型

5.3.2.1 异常数据滤波模型和主成分分析去噪模型

去除噪声数据是数据预处理的一项基本处理过程。针对从基因表达谱这类特殊的数据中发现基因标签这类特点问题，我们讨论一般的滤波去噪方法并提出一种主成分分析的去噪模型。

考虑到基因表达谱中的数据主要是来源于在癌症患者和正常人的基因上通过微观实验和生物化学原理等复杂技术获取到的。鉴于微观实验操作难度，系统误差可能导致突兀数据。同时考虑到癌症这种本身就是生理及其异常的疾病，即使是患有同一种癌症的病人也可能存在某些局限于个别人的基因突变的情况。考虑到研究中主要是发现问题的一般规律，我们做如下假设：

1. 某些癌症病人的基因突变引发的数据突变视为异常数据
2. 由于数据具有高噪声、波动大的特点，将系统误差和疏忽误差引起的

突兀数据作为异常数据

3. 数据中一定存在噪声，主要是由测量的随机误差和固有的系统的误差共同决定，并且噪声数据符合高斯分布。

1）异常数据滤波模型

我们建立去除异常数据带来的噪声的模型。主要采用两种方法：均值滤波和中值滤波。通过观察，我们发现基因表达谱数据中存在突兀的数据项，即基因的样本数据严重偏离该样本的均值，下面统一称为异常数据。我们分别使用均值和中位数作为数据的参考基准，实验中设定阈值并通过调整其大小来设定过滤异常数据的滤波窗口的大小。

均值滤波算法过程如下：

Step 1. 设定滤波窗口阈值 a 和调整数据比例阈值 b=5%；

Step 2. 对每个正常人和癌症病人分别计算 Step3；

Step 3. 计算每个基因维度数据的均值，并根据窗口阈值调整落在窗口外面的奇异的到窗口边缘；

Step 4. 统计被调整的奇异数据项的数量，并计算其在整个数据中的比例。调整比例如果约为 5%则终止程序，否则跳转到 Step 1 按照一定步长调整 a 的值。

中位数滤波算法类似，不再赘述。下面主要从基因谱数据挖掘基因标签这个具体应用出发构建主成分分析去噪模型。

2）主成分分析去噪模型

主成分分析是一种采用组合特征的方法将多维数据降维的方法。方法的目标是寻找在最小化重构误差的意义下最能够代表原始数据的投影方法。降维后的数据能够比较好的代表原始数据。主成分分析的主要思想是：

1. 用一维向量表示高维样本
2. 将一维投影量扩展到相对低维的空间
3. 低维空间是由高维空间数据的散布矩阵的最大几个特征值向量构成

通常情况下高维空间数据的散步矩阵的最大几个特征值占据了特征之和的绝大部分，所以可以认为少数几个最大特征值对应的特证向量即可表示原数据中的绝大部分信息，而剩下的小部分（即对应较小的特征值的特征向量所表示的信息），通常可以认为是数据噪声而丢掉。考虑到基因谱数据的高噪音、多异常、大波动的特点，我们可以通过主成分分析的方法在降低维度的过程中去除噪音数据。

主成分分析一般是对样本的特征属性维度进行降维，在基因谱数据中即对应于基因维度。一方面由于主成分分析降维中是将当前的维度空间映射到低维空间，映射后将会当前维的多个维度映射到低维空间某个维度上，从而降维后的数据维度并不对应于某个当前维度；另一方面考虑到基因的维度在选取特征基因问题要求下需要是不能在降维去噪的过程中被“坏掉”的。所以我们考虑从样本的维度使用主成分分析方法进行降维去噪。

我们将数据集合按照样本的种类分成多个数据集合（这里是两个类别）分别进行主成分分析，这样降维后的每个数据项可以看成一个新的样本，并且其类别保持降维前所属的类别。

5.3.2.2 随机噪音的高斯模型

由随机误差和固有的系统误差引入的噪音有很好的随机性，一般假设其符合

某种自然分布，其中以高斯分布最为普遍。下面介绍两种使用概率方法基于噪音模型的建模方法。

下面首先介绍一种对噪音建模的高斯过程分类器（Gaussian Process Classification）[6]，其基于贝叶斯（Bayes）理论的概率学习算法，使用高斯过程模型对噪音建立模型效果很好。

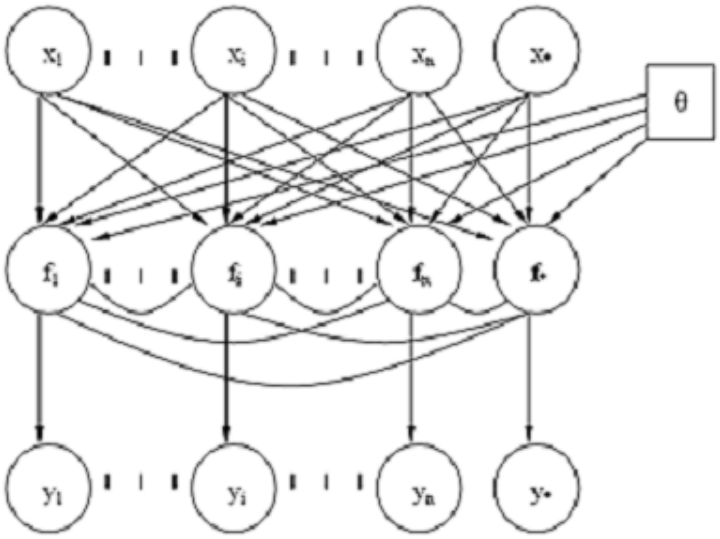

图 5.3.1 GPC 图模型示意图[6]

高斯过程分类器（简称为 GPC）可以看作是一个图模型（如图 5.3.1），用随机变量表示输入、潜在变量表示函数值和类别标签。潜在函数值完全决定类别标签。有很多噪音模型用来建模类别标签的似然函数。该问题可以如下作形式化定义：

仅考虑二分类问题，假定有数据集 $D=\{(x_i,y_i)\,|\,i=1,2,...,n\}$ 其中 x_i 表示实例，$y_i\in\{0,1\}$ 表示两类标签。在训练数据集上，我们希望训练得到一个分类模型能通过计算新实例 x_* 的可能性最大的所属类别 $p(y_*\,|\,x_*,D)$。

高斯过程分类器的核心思想是假设可以通过一些与 x_i 关联的且真正有价值的隐藏变量 $f(x_i)$ 来判别出实例的类别 y_i。下面阐述从贝叶斯框架的角度建立 GPC 模型的主要步骤。

首先，我们在函数 $f(\cdot)$ 上设定一个先验概率，即给定一个有限集合 $X=\{x_1,...,x_m\}$，随机向量 $f=[f(x_1),...,f(x_m)]^T$ 服从高斯分布。不是一般性我们假设过程的期望为 0，且有 $f(x_i)$ 和 $f(x_j)$ 的协方差为：

$$\Sigma_{ij} = c(x_i, x_j) = v_0 \exp\{-\frac{1}{2}\sum_{m=1}^{h} l_m (x_i^m - x_j^m\} + v_1 + v_2\delta(i,j)$$

其中超参数v_0反映了隐藏变量方差的垂直波动，v_1表示隐藏变量离 0 期望的偏置，v_2表示隐藏噪音的方差（δ表示噪音，期望为 0，方差为σ^2），l_m表示第 m 个特征属性在模型中的贡献权重。隐藏函数值$\{f(x_i)\}$服从多元高斯分布：

$$p(f) = \frac{1}{(2\pi)^{\frac{n}{2}} \Sigma^{\frac{1}{2}}} \exp(-\frac{1}{2} f^T \Sigma^{-1} f)$$

其次，确定似然值为给定样本和隐藏函数值的结合，是似然函数的乘积：

$$p(D \mid f) = \{\prod_{i=1}^{n} p(y_i \mid f_i)\}$$

我们假定隐藏函数值被高斯噪音影响，并且和输入独立。其中考虑了高斯噪音后的似然函数为：

$$p(y_i \mid f_i) = \Phi(\frac{y_i f_i}{\sigma})$$

第三，可以得到后验概率：

$$p(f \mid D) = \frac{1}{p(D)} \prod_{i=1}^{n} p(y_i \mid f_i) p(f)$$

第四，根据文献[6]中的推导可以预测给定的实例x_*的类别y_*的分布：

$$p(y_* \mid x_*, D, \theta^*) = \int p(y_* \mid f(x_*), \theta^*) p(f(x_*) \mid D, \theta^*) df(x_*)$$

其中θ^*是假设发现的最优超参数。

至此建立了贝叶斯框架下的基于噪声模型的高斯过程分类器。

文献[6]中使用 Expectation Propagation 算法求解 GPC 模型，并在 colon cancer 数据集（和本文实验的数据集十分类似）上面进行试验，得到如下结果：

表 5.3.1 2000 个基因的测试错误率

Algorithms	Test Error Rate (%)	Predictive Variance
SVM	0.226	-
Laplace-MAP	0.219	3.49
EP	0.19	6.72

从这个结果可以看出基于噪音模型的 GPC 模型可以取得比较好的结果。

5.3.3 问题三的讨论

从上面的论述可以看出，噪音在建立优秀的分类器，尤其是基于概率模型的分离器中起到了很重要的角色。下面我们将从分类器的泛化能力的角度，阐述一下噪音扮演的重要作用。

我们在确定基因标签的过程中主要是要判别基因标签对癌症的识别能力，通

常的方法是使用基因组合作为特征属性训练出一个分类器，并通过分类器的分类能力来鉴别基因组合是否为基因标签（参考第一、第二问的解决方案）。

而数据中存在噪音是不可避免的，同时也是有其优点的。在使用机器学习的方法训练一个分类器的过程中，我们需要避免的一个问题就是分类器对训练数据的过渡拟合。

过渡拟合：是训练获得的模型过于符合训练数据的特点，而泛化能力较弱，从而在未知的新的测试数据集上面的分类效果一般，甚至错误较多。

这样噪音的存在可以再很大程度上扰动了分类器的容忍能力，使他不可能完全拟合训练数据，从而对新的测试样本有比较好的识别能力。而且在确定肿瘤的基因标签的背景之下，未知类别的测试样本的数据异常情况比较多见，这时候分类器的泛化能力显得尤为重要。所以基于噪音较大的数据发现出来的基因标签的对这些异常癌症样本的识别能力在一定程度上得到保证。

5.4 问题四的分析、建模与求解

5.4.1 问题四的分析

问题一、二、三的基因图谱分析模型都是以基因图谱的统计数据为基础进行分析的，而基因图谱信息不可避免的含有噪声，而这些噪声会影响特征肿瘤信息基因的确定。实际在肿瘤的研究领域，根据临床经验会已知若干个基因与某种癌症的关系密切，因此将包含临床经验的知识库融入到基因图谱模型中更加有利于癌症信息基因的确定。

题目中已知信息临床生理学信息：大约有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。根据这个信息可以建立结肠癌的知识库，在基因表达谱的分析上可以重视知识库中的基因，提高其重要性参数，因此得到的肿瘤信息标签在具有很好的结肠癌判别能力的同时，更加尊重了实际临床数据的重要性。

该部分首先分析我们提出的基于知识库的基因图谱分析模型（Knowledge-based FSSM VSM ，KFS模型），然后结合题目中给定的结肠癌数据计算结肠癌的信息基因，并分析KFS模型与问题二的基因图谱分析模型间处理结果的分类能力比较，该部分最后给出结果分析以及出现该结果的原因，并提出了本模型还需要解决的问题。

5.4.2 问题四的模型建立

基于上述对问题四的分析，本文提出一种基于知识库的基因图谱分析模型(KFS模型)，该算法引入信息基因知识库的概念，模型首先基于临床经验建立模型的知识库，对去噪后的数据进行样本评价函数增益，然后对剔除基因图谱中的无关基因，降维后的数据通过KFSSM（Knowledge-based FSSM）算法的处理得到分类特征子集空间，然后模型分别采用改进的支持向量机（ISVM）以及人工神经网络（IANN）对分类特征子集的分类能力进行考察，最终确定癌症的信息基因组合，KFS模型的结构图如图5.4.1所示。

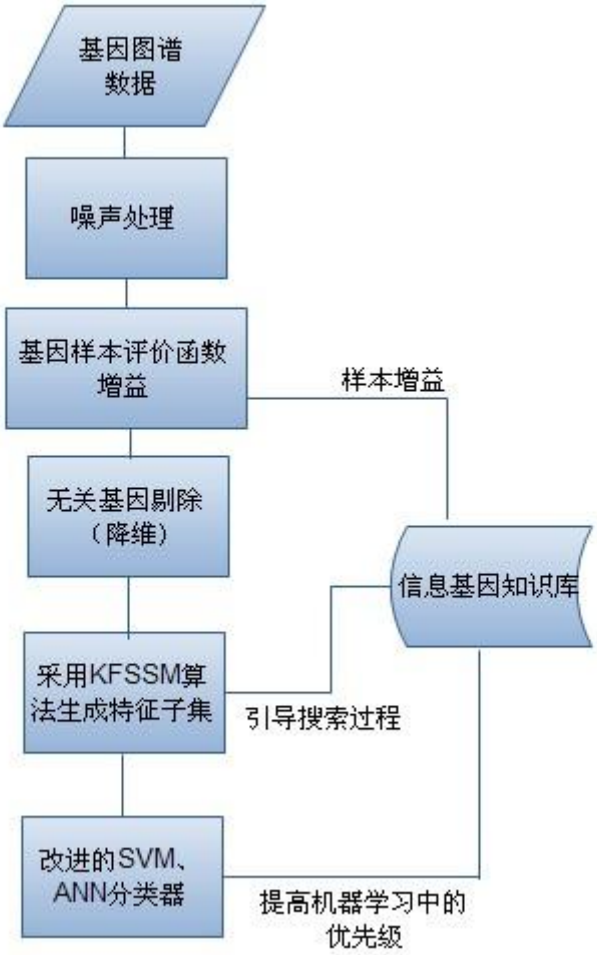

图 5.4.1 基于知识库的基因图谱分析模型

5.4.2.1 癌症基因知识库

题目中提到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，那么在基因图谱分析的中引入已知关系必然会提高肿瘤信息基因判别的实际准确率和有用性。本文提出一种肿瘤信息基因知识库概念，用于存储临床经验已经发现的肿瘤信息基因。

结合题意，本文提出的信息基因知识库具有可以表示如下：

[基因名称， 基因变化，基因样本统计概率] （5.1）

其中，基因名称表示与该类肿瘤的信息基因名称（本文算法中采用编号标识）；基因变化分为突变和失活（本文算法中取值分别为1、-1），基因突变表示基因表达水平值明显上调，基因失活表示基因表达水平值明显下降，表现在基因图谱中是基因的表达水平出现较大波动；基因样本统计概念表示该基因在样本统计中发生变化（突变或失活）的概率。

假设知识库的一项记录为[g_i , 1 , k%]，参考表达式5.1可知该项记录意思是信息基因g_i在临床经验中有k%的样本表现出基因突变特征。

引入信息基因知识库的概念就是为了帮助从基因信息图谱中判断出肿瘤的信息基因组，考虑信息基因与癌症的密切关系，本文提出的信息基因知识库在KFS模型中的作用主要有三个：

1）知识库在数据预处理中的作用

由于基因表达谱中不可避免的存在噪声等因素的干扰，并且由于基因图谱中样本数目相对于基因数目往往很少，那么即使是信息基因也可能因为噪声数据而没有被选为信息基因，因此需要知识库来修正这个结果。

由于知识库样本数据的多样性以及临床数据的可供参考性，参考知识库中的信息基因修正相应的样本值：通过增加该基因的评价参数（巴氏距离），则可以使得该基因表现的更像信息基因。

2）知识库在分类特征子集生成过程中的作用

FSSM算法的特征子集空间搜索过程中的评价函数是以Fi的Bhattacharyya距离为评价函数，这忽视了知识库中信息基因的重要性，在比较selectMax[i]与Fmax[i]的更新问题上就需要判断是否需要更新，具体算法见KFSSM算法的描述。

需要说明的是，虽然信息基因知识库中基因具有更重要的作用，但是不会出现在所有的特征子集中，这也是符合算法执行的情况的，具体分析见问题四的结果及分析部分。

3）知识库在分类器中的作用

本文同样将知识库的作用考虑进了支持向量机以及人工神经网络分类器中，在机器学习的过程中考虑知识库中基因的重要性，提升了知识库中基因在学习过程中的重要性。

综上所述，知识库的本质意义在于对基于基因图谱数据的处理过程中起到一个引导作用，使得结果在某种程度上偏向于具有较高参考价值的信息基因知识库。

5.4.2.2 基因图谱数据预处理

由于一般基因图谱中，样本相对于基因数目往往很少，如果直接用于分类会造成小样本的学习问题，因此需要对基因图谱的原始数据进行预处理，本文KFS模型中的图谱数据预处理主要包含三个功能：噪声处理、基于知识库的样本评价函数增益和无关基因剔除，该部分试图通过对原始基因图谱数据进行分析得到初步过滤的特征子集，其结构如下图5.4.2所示。

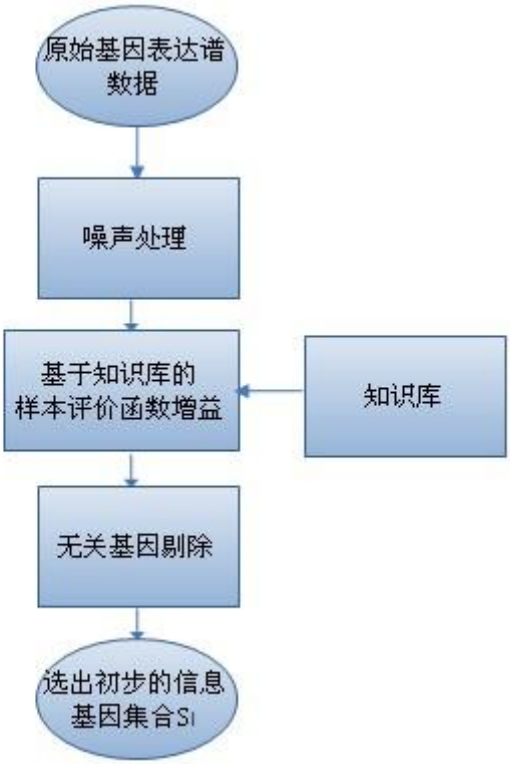

图 5.4.2 KFS模型中预处理结构图

1）噪声去除

由于在读取生物芯片时一些不可控因素导致了某些基因样本表达水平发生了很大的变化，因此信息基因的信号有可能被噪声淹没，因此在基因图谱数据分析前需要进行噪声处理。

本文提出的KFS模型中的噪声去除直接采用了问题三建立的噪声模型，去除了基因表达谱中的噪声数据，为后续的基因图谱分析提供较好的数据。

2）基于知识库的样本评价函数增益

本文采用Bhattacharyya距离作为衡量基因中蕴含的分类信息量的评价函数，基因的Bhattacharyya距离越大，该基因对于样本的分类能力就越强，其被选为信息基因的可能性就越大。

基于信息基因知识库的分析，知识库中的信息基因对判断正常样本和癌症样本的区分能力很强，体现在样本的评价函数上就是需要大的Bhattacharyya距离，因此本文对基因的Bhattacharyya距离做简单的增益，以增加基因图谱数据分析过程中该基因被选为信息基因的可能性。

假设[g_i, x, k%]为知识库中的一个信息基因记录，增益后的评价函数为：

$$Bhattacharyya(g_i) = Bhattacharyya(g_i) + \alpha * (k\% * maxB)$$

其中，增益参数经α的选择取决于实际基因图谱数据,maxB表示样本中最大的Bhattacharyya距离值。经过该增益后，知识库中的基因在基因图谱中的相对表达水平获得一定的增益，增益的程度取决于知识库中基因样本统计概率。

因此在通过设定Bhattacharyya距离阈值θ来剔除冗余基因的过程中，由于知识库中的基因在选择过程中获得了相应的增益，所以其被选进信息基因集合S_I

的可能性就增加了。

3）无关基因剔除

本文的KFS模型中的无关基因的剔除采用了Bhattacharyya距离，在上一节增益的基础上，通过设定阈值 θ 来生成无关基因集合S_N和信息基因集合S_I。

5.4.2.3 基于知识库的 FSSM 的分类特征子集生成算法

通过KFS模型预处理过程，我们得到了一组约为样本规模10%的信息基因集合S_I，本文提出一种基于知识库的FSSM算法（KFSSM）算法，对基因集合S_I构成的特征子集空间进行搜索，最后得出34组具有不同维数的分类特征子集。

本文提出的KFSSM算法的动态搜索过程中，也采用了Bhattacharyya距离作为评价函数，同时基于知识库引导KFSSM算法的搜索过程。KFSSM算法的基本原理是在顺序搜索的过程中，当计算出由i+1个基因的构成的最大评价函数的基因集合Fmax[i+1]时，回溯的时候计算出其中i个基因构成的最大评价函数SelectMax[i]，此时在刷新Fmax[i]的时候需要考虑知识库中的信息基因，对于包含知识库中信息基因权重较大的特征子集最后被选中的可能性较大，特征子集在知识库中的权重函数的设定设定如下：

$$Weight(F_i) = \sum_{g_i \in KDB} k(g_i)$$

其中，KDB表示信息基因知识库，k表示基因的样本统计概率，g_i表示特征子集F_i中包含的知识库中的基因。

算法评价的基本流程如图5.4.3所示，其中Fmax[i]表示含有i个基因的特征子集的且具有最大评价函数的基因组合，selectMax[i]表示从当期i+1个基因中选择i个取其评价函数最大的特征子集，B(Fmax[i])函数表示求特征子集的Bhattacharyya距离，W()函数表示求相应特征向量在知识库中的权重值。

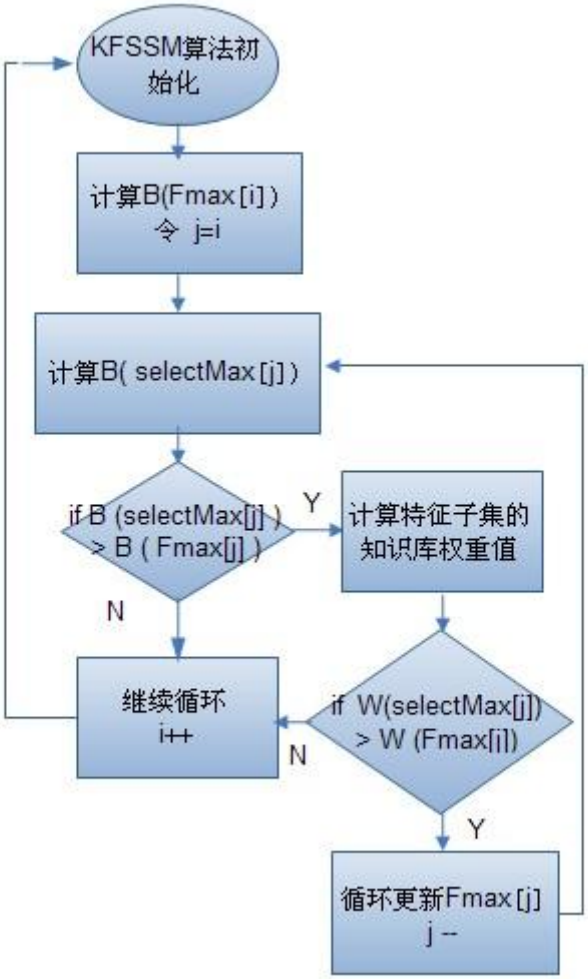

图 5.4.3 KFSSM算法搜索过程原理图

本文采用matlab程序对KFSSM算法进行了实现，下面列出算法的伪代码如下：

```
设置分类特征子集空间大小target值 %target = 35
读取信息基因集合SI和知识库数据kdbArray
初始化Fmax[2]数据    %直接取B值最大的两个向量

while i < target
    SI =SI - Fmax[i]; %从候选基因中删除已选基因
    计算Fmax[i+1] = max{Fmax[i],g} ; %选择Fmax[i]与SI中基因拥有最大评价函数的组合

    j = i;  %用于控制回溯
    while true
        计算selectMax[j];  %从Fmax[i+1]中选择i个基因组合的具有最大评价函数的组合

        if B( selectMax[j] ) > B( Fmax[j] )  &&   W(selectMax[j]) > W(Fmax[j])
            Fmax[j] = selectMax[j] ;  %更新Fmax[j]值

            if j==2
```

```
                    break;     %回溯更新执行到Fmax[2]停止
                else
                    j = j - 1;
                end
            else
                break;      % 不需要回溯更新
            end
        end
    end
end
循环结束，获得target组不同维数的分类特征子集
```

5.4.2.4 改进型的 SVM 和 ANN 分类器

本文的KFS模型中的支持向量机和人工神经网络方法使用问题二的模型，本文同样将知识库的作用考虑进了支持向量机以及人工神经网络分类器中，并对其做了简要修改，在学习的过程中考虑知识库中基因的重要性，提升了知识库中基因在学习过程中的重要性。

知识库在支持向量机和人工神经网络的运用是本文需要改进的地方，后续工作是实现基于知识库的分类器模型。

5.4.3 问题四的模型求解

文章结合给出的基因图谱信息以及结肠癌的临床数据，利用本文提出的KFS算法处理基因表达谱数据，下面描述处理过程。

5.4.3.1 结肠癌信息基因知识库的建立

题目中指出临床经验：大约90%结肠癌在早期有5号染色体长臂APC基因的失活，40%~50的ras相关基因突变，因此根据这些已知关系可以建立结肠癌信息的知识库。

搜索文章中给出的基因图谱找出APC基因的位置是project_data1.xls文件第869行的Hsa.2238基因，经过去除重复基因得到该基因在本文算法中的编号823，因此可得关于 Hsa.2238基因的记录：[Hsa.2238 ， -1 ， 90%]; 搜索ras相关基因可得其两个记录分别为Hsa.2974和Hsa.2868 ，它们在本文算法中对应的编号分别为1027和1044，因此记录分别为：[Hsa.2974 , 1, 40%~50%]和[Hsa.2868, -1, 40%~ 50%]。

为方便本文算法计算，文章中的基因名称替换成在本文算法中的基因编号，因此根据题意，建立起关于结肠癌信息基因的知识库如表格5.4.3.1所示。

表 5.4.1 已知结肠癌信息基因知识库

[823 ， -1 ， 90%]
[1027 , 1, 40%~50%]
[1044, -1, 40%~ 50%]

5.4.3.2 采用 KFS 模型处理结肠癌图谱分析过程

1）样本评价函数增益

首先对结肠癌基因表达谱数据进行去除噪声处理，然后根据本文KFS算法对于知识库中信息基因的增益，选择增益参数 α = 1，算法执行的结果如表格

5.4.3.2所示。

表格5.4.2 增益前后基因知识库中基因的选择情况

	基因编号	Bhattacharyya距离	选择情况
增益前	823	0.0778264	都被当成无关基因剔除
	1027	0.00152048	
	1044	0.0179445	
增益后	823	0.401826	全部被选择
	1027	0.16352	
	1044	0.179945	

经过样本评价函数增益后，增益前被当成无关基因剔除的第823、1027和1044号基因加入到信息基因集合S_I中，分析过程的下一步骤采用KFSSM算法对信息基因集合进行动态搜索生成具有不同维数的分类特征子集。

2）基于KFSSM算法的特征子集生成

经过样本评价函数增益后的得到的初步信息基因集合S_I，本文KFSSM算法的作用就是在该部分的基础上生成特征子集。

本文KFSSM算法的执行结果参考附录2，由数据可以看出：第823号基因被全部的特征子集包含，而所有的特征子集都没有包含第1027和第1044号基因，关于算法执行的结果将在问题四的结果及分析中进行详细的分析。

3）SVM分类器的选择结果

本文算法对上一步生成的34组具有不同维数的特征子集进行考察，选出具有最佳分类能力的特征子集（即信息基因的组合），算法执行的结果是：（Hsa.2238，Hsa.2800，Hsa.5392，Hsa.1454，Hsa.2291，Hsa.2487，Hsa.3331，Hsa.43331，Hsa.40063，Hsa.8214）具有最好的样本分类能力，样本分类正确率达到94.52%，其中第823号基因名称为Hsa.2238。

表格5.4.3.3列出基于支持向量机的处理的分类正确率最好的前5个基因组合，基于人工神经网络的样本识别结果见附录3。

表5.4.3 基于支持向量机的分类正确率最好的前5个基因组合

样本分类正确率	基因组合
94.52%	(Hsa.2238，Hsa.2800，Hsa.5392，Hsa.1454，Hsa.2291，Hsa.2487，Hsa.3331，Hsa.43331，Hsa.40063，Hsa.8214 ）
93.37%	(Hsa.2238, Hsa.2800,Hsa.5392, Hsa.1454, Hsa.2291,Hsa.2487, Hsa.3331,Hsa.43331,Hsa.40063, Hsa.8214, Hsa.6288 Hsa.1165, Hsa.2842, Hsa.6317)
92.79%	(Hsa.2238, Hsa.5392, Hsa.1454, Hsa.2291, Hsa.2487,Hsa.3331,Hsa.43331, Hsa.40063)
92.55%	(Hsa.2238, Hsa.2800, Hsa.5392, Hsa.1454, Hsa.2291, Hsa.2487, Hsa.3331, Hsa.43331, Hsa.40063, Hsa.8214, Hsa.6288,Hsa.1165)
91.76%	(Hsa.2238, Hsa.2800, Hsa.5392,Hsa.2291,Hsa.2487, Hsa.3331,Hsa.43331, Hsa.40063, Hsa.8214, Hsa.6288, Hsa.1165)

5.4.4 问题四的结果及分析

5.4.4.1 处理结果及分析

本文提出了基于知识库的基因图谱分析模型，依据临床经验建立的知识库各部分的处理结果如下：

1）样本评价函数增益结果

原来被当成无用基因剔除的第823、1027和1044号基因加入到信息基因结合S_I中。

分析其原因可得：无关基因的剔除采用了Bhattacharyya距离阈值剔除的方法，由于样本数目过少及噪声都可能导致知识库中的信息基因的Bhattacharyya距离值偏小从而被提出，通过对知识库中的信息基因的Bhattacharyya距离增益，可以避免有用基因在剔除无关基因的过程中就被剔除。

2）KFSSM算法执行结果

第823号基因被选为信息基因（参考附录2），而第1027和1044号基因没有被选入信息基因集合S_I。

分析其原因是KFSSM算法执行过程中特征子集的生成依赖于特征子集的Bhattacharyya距离，若算法执行中设置的增益参数较小，则有可能导致知识库中的某信息基因没有被选进特征子集空间。需要注意的是，实际实验过程中可以通过调高样本评价函数增益参数来提高知识库中样本对基因图谱分析的影响力。

3）KFS模型执行结果与问题二模型执行结果比较

本文提出的基于知识库的基因图谱分析模型执行的结果是：基因组合（Hsa.2238，Hsa.2800，Hsa.5392，Hsa.1454，Hsa.2291，Hsa.2487，Hsa.3331，Hsa.43331，Hsa.40063，Hsa.8214）具有最好的样本分类能力，样本分类正确率达到94.52%；第二问执行的结果是：基因组合（Hsa.37937，Hsa.710，Hsa.3016，Hsa.5392，Hsa.6080），样本分类正确率达到95.79%。

5.4.4.2 结论分析

1、基于知识库的基因图谱分析结果包含知识库中的基因

KFS模型执行的结果中包含知识库中的基因Hsa.2238，执行结果在保证样本分类正确率的基础上尊重临床经验，显然会对结肠癌分类研究提供更大的帮助。知识库中的Hsa.2974和Hsa.2868基因没有出现在最终基因组合中，原因是上述基因在知识库中的样本统计概率值较低，且由于数据噪声的原因，导致最终的信息基因没有包含它们。

2、样本分类正确率比较

同样以支持向量机为分类器，KFS模型的最好的样本分类正确率为94.52%，而问题二中的样本分类正确率为95.79%，KFS模型出现分类正确率下降是因为为了考虑知识库中的信息基因，KFS模型通过样本评价函数增益及KFSSM算法使得最终的结果朝着包含知识库基因的方向发展，这在某种程度上忽略了基因图谱原始数据在分类器中的作用。

3、两者的共同基因及分析

观察发现KFS模型计算出的信息基因组合与问题二中计算出的信息基因都包

含Hsa.5392基因，这是因为Hsa.5392基因具有较大的Bhattacharyya距离，因此包含更多的分类信息量，基因Hsa.5392的数据分布如图5.4.4.1 所示。

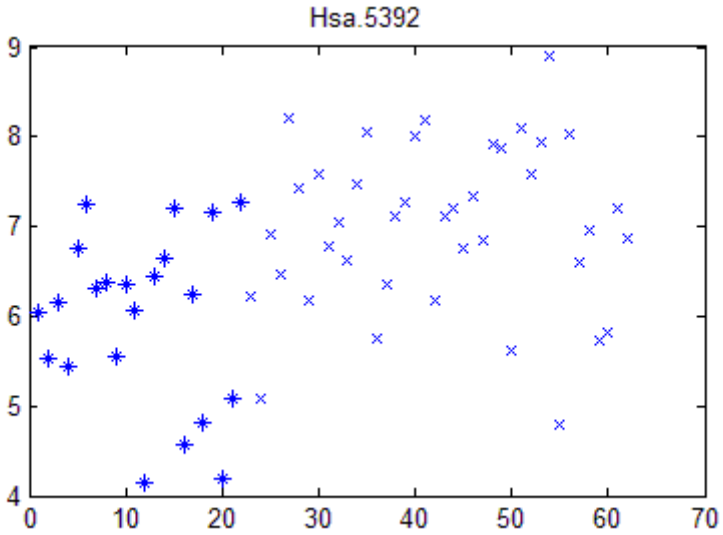

图5.4.4 基因Has.5392样本数据分布

4、KFS模型还需解决的问题

KFS模型的分类器还需要进一步修改，在结果中引入知识库因素，这样得到的基因组合不但能够考虑到知识库中的信息基因，还能够提高基因组合的样本分类正确率。

六、总结

6.1 模型的总结

（1）问题一采用简单的巴氏距离和相似度模型，达到了很好的分类效果，剔除了大量的冗余基因；

（2）问题二的 FSSM 算法有效地在信息基因空间搜索候选特征子集，通过评价函数 J 函数的设定，选出具有不同维数的分类特征子集，然后分别采用 SVM 和 ANN 分类器进行样本识别，最终获得分类正确率最好的信息基因集合；

（3）问题三阐述了噪音在本题的两种角色，一种是作为异常数据对之建立去噪模型，另一种是作为随机因素对之建立容噪模型。同时说明了噪音可以防止分类器过渡拟合而获得较好的泛化能力；

（4）本文问题四中提出一种基于知识库的基因图谱分析模型 KFS 模型，首先对基因图谱数据进行去噪、评价函数增益以及剔除无关基因，然后采用提出的 KFSSM 算法生成分类特征子集，模型最后采用改进的 SVM 和 ANN 算法计算出最大分类正确率的信息基因组合。

参考文献

[1] Golub TR, Slonim DK, Tamayo P. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J] . Science ,1999 ,286(5439) :531 - 537.

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

[3] 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, pp.240-244(2005).

[4] Wenlong Xu, Xianghua Zhang, Huanqing Feng. Using Simple Gaussian Mixture Model for Multiclass Classification Based on Tumor Gene Expression Data. ICBBE2008. Shanghai, China, May 16-18.

[5] 贾民平,张洪亭,周剑英. 测试技术. 高等教育出版社，2001.12.

[6] Mingyue Tan. Expectation Propagation of Gaussian Process Classification and Its Application to Gene Expression Analysis.

[7] 李颖新, 朱云华, 阮晓钢. 基于支持向量机的肿瘤亚型判断. 博士论坛.

[8] 王守觉, 周凌飞. 基因表达数据分析中的特征基因提取, 软件时空, Vol 24.

附录 1：FSSM 法生成的 34 个特征子集

特征子集大小	特征子集
2	Hsa.37937 Hsa.5392
3	Hsa.37937 Hsa.710 Hsa.5392
4	Hsa.37937 Hsa.710 Hsa.3016 Hsa.5392
5	Hsa.37937 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080
6	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080
7	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058
8	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331
9	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.43331 Hsa.8214 Hsa.823
10	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957
11	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965
12	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965
13	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.823 Hsa.33965 Hsa.816 Hsa.490 Hsa.732
14	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.823 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689
15	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689
16	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928
17	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147
18	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147
19	Hsa.37937 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250
20	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250
21	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048
22	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732

	Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58
23	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283
24	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572
25	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186
26	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221
27	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562
28	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562 Hsa.1140
29	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562 Hsa.1140 Hsa.8125
30	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562 Hsa.1140 Hsa.8125 Hsa.94 Hsa.56
31	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562 Hsa.1140 Hsa.8125 Hsa.94 Hsa.56
32	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562 Hsa.1140 Hsa.8125 Hsa.94 Hsa.56 Hsa.1610
33	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058

	Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562 Hsa.1140 Hsa.8125 Hsa.94 Hsa.56 Hsa.1610 Hsa.10176
34	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562 Hsa.1140 Hsa.8125 Hsa.94 Hsa.56 Hsa.1610 Hsa.10176 Hsa.285
35	Hsa.37937 Hsa.549 Hsa.710 Hsa.3016 Hsa.5392 Hsa.6080 Hsa.2058 Hsa.43331 Hsa.8214 Hsa.823 Hsa.957 Hsa.33965 Hsa.816 Hsa.490 Hsa.732 Hsa.36689 Hsa.2928 Hsa.8147 Hsa.6814 Hsa.2250 Hsa.7048 Hsa.58 Hsa.41283 Hsa.33572 Hsa.42186 Hsa.1221 Hsa.562 Hsa.1140 Hsa.8125 Hsa.94 Hsa.56 Hsa.1610 Hsa.10176 Hsa.285 Hsa.695

附录 2　基于 KFS 模型的 KFSSM 算法生成的特征子集空间

上述算法生成了 34 个具有不同维数的特征子集向量，其中特征子集中的基因以编号表示，如下图表格所示。

表格 6.2 KFSSM 算法执行结果

	分类特征子集
1	823 463
2	823 463 64
3	823 463 64 1673
4	823 463 64 1673 1281
5	823 463 64 1673 1281 1721
6	823 463 1673 1281 1721 1261 1595
7	823 1281 1595 1760 542 915 1838 1016
8	823 1673 1281 1595 1760 542 915 1838 1016
9	823 64 1281 1595 1760 542 915 1838 1016 1160
10	823 64 1281 1760 542 915 1838 1016 1160 1377 531
11	823 64 1281 1595 1760 542 915 1838 1016 1160 1377 531
12	823 64 1281 1595 1760 542 915 1838 1016 1160 1377 531 1264
13	823 64 1281 1595 1760 542 915 1838 1016 1160 1377 531 1684 1264
14	823 64 1281 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264
15	823 64 1281 1595 1760 542 915 1838 1016 1160 1377 531 1789 1264 1877 930
16	823 64 1281 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930
17	823 64 1673 1281 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930
18	823 64 1673 1281 1261 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930
19	823 64 1673 1281 1595 1760 542 915 1838 1016 1160 1377 531 1789 1264 1877 930 308 1477 777
20	823 64 1673 1281 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930

	308 1477 777
21	823 463 64 1673 1281 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 1477 777
22	823 463 64 1673 1281 1261 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 1477 777
23	823 463 64 1673 1281 1261 1595 1760 915 1838 1016 1160 1377 531 1789 1264 1877 930 308 230 1477 777 414 661
24	823 463 64 1673 1281 1595 1760 915 1838 1016 1160 1377 531 1789 1264 1877 930 308 230 1477 777 414 661 880 612
25	823 463 64 1673 1281 1261 1595 1760 915 1838 1016 1160 1377 531 1789 1264 1877 930 308 230 1477 777 414 661 880 612
26	823 463 64 1673 1281 1261 1595 1760 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1477 777 414 661 880 612
27	823 463 64 1673 1281 1261 1595 1760 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1477 777 414 661 880 612 1200
28	823 463 64 1673 1281 1261 1595 1760 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1477 777 414 661 880 612 1200 583
29	823 463 64 1673 1281 1261 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1477 777 414 661 880 612 1200 583
30	823 463 64 1673 1281 1261 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1477 777 414 661 880 612 1200 583 1577
31	823 463 64 1673 1281 1261 1595 1760 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1495 1477 777 414 661 880 612 1200 583 1577 586
32	823 463 64 1673 1281 1261 1595 1760 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1495 1477 777 414 661 880 612 1200 583 1577 586 156
33	823 463 64 1673 1281 1261 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1495 1477 777 414 661 880 612 1200 583 1577 586 156
34	823 463 64 1673 1281 1721 1261 1595 1760 542 915 1838 1016 1160 1377 531 1684 1789 1264 1877 930 308 230 1495 1477 777 414 661 880 612 1200 583 1577 586 156

从KFSSM数据可以看出：第823号基因被全部的特征子集包含，而所有的特征子集都没有包含第1027和第1044号基因。

附录 3 基于人工神经网络的分类正确率最好的前 5 个基因组合

样本分类正确率	基因组合
98.39%	(Hsa.2238, Hsa.2800, Hsa.5392, Hsa.1454, Hsa.2291, Hsa.2487, Hsa.3331, Hsa.43331, Hsa.40063 , Hsa.8214)
96.77%	(Hsa.2238, Hsa.5392, Hsa.1454, Hsa.2291, Hsa.2487, Hsa.3331, Hsa.43331, Hsa.40063)
95.16%	(Hsa.2238, Hsa.2800, Hsa.5392, Hsa.2291, Hsa.2487, Hsa.3331, Hsa.43331, Hsa.40063, Hsa.8214, Hsa.6288, Hsa.1165)
95.16%	(Hsa.2238, Hsa.2800, Hsa.5392, Hsa.1454, Hsa.2291, Hsa.2487,Hsa.3331 Hsa.43331, Hsa.40063, Hsa.8214, Hsa.6288, Hsa.1165)
95.16%	(Hsa.2238, Hsa.2800, Hsa.5392, Hsa.1454, Hsa.2291, Hsa.2487,

	Hsa.3331, Hsa.43331, Hsa.40063, Hsa.8214, Hsa.6288, Hsa.1165, Hsa.2842 Hsa.6317)

附录 4 源程序文件功能介绍

I．C++程序说明

主要功能说明：计算巴式距离、相似度以及信息基因集 S_I

程序文件	功能	输入输出
计算巴氏距离.exe	计算巴氏距离	输入文件： matrix_geneRow.txt
计算相似度.exe	计算基因相似度	输入文件：guiyi.txt 输出文件： similityFile.txt
FirstSelect.exe	计算信息基因集	输入文件： firstselectBPart.txt firstselectSPart.txt 输出文件： firstselect.txt

II．Matlab 程序说明

主要功能说明：生成不同维数的分类特征子集

程序文件	功能	输入输出
FSSM.m	读取剔除无关基因后的信息基因，生成不同维数的分类特征子集	输入文件:firstselect.txt data_matrix.txt 输出文件: secondselect.txt
KFSSM.m	基于知识库的FSSM算法:读取信息基因，生成不同维数的分类特征子集	输入文件：kdb.txt firstselectKFSSM.txt data_matrix.txt 输出文件： secondselectKFSSM.txt

III．Java 程序说明

主要功能说明：提供包含支持向量机及人工神经网络的分类器功能

问题说明	函数使用	函数说明
使用机器学习算法构建基于特征基因组合的分离器来评估特征基因的分类能力，进而选择基因标签	seu.edu.cn.learner.attrselect .AttributeSetSelector	评估候选特征基因组合的分类能力，排序输出
	seu.edu.cn.data.AttributeSelector	根据特征基因组合构造学习实例集

		合
	seu.edu.cn.learner.NeuroNetworkLearner	基于 Weka 的人工神经网络学习算法
	seu.edu.cn.learner.SVMLearner	基于 Weka 的支持向量机学习算法

使用说明：

1. 建立 java 工程，导入 jar 包 weka.jar，libsvm.jar，wlsvm.jar
2. 导入数据文件
   a) data_matrix2_2.txt　包含所有的预处理后的基因表达谱数据
   b) data_set_2.txt　包含待评估的基因组合，一行一组，组内使用空格分隔
3. 运行 AttributeSetSelector 主程序，输出按照分类能力由大到小排好序的特征基因组合