多类别肿瘤基因表达谱的自动特征选择方法

高 娟 王国胤 胡 峰

(重庆邮电大学计算机科学与技术学院 重庆 400065)

摘 要 从信息学角度出发寻找肿瘤相关基因、发现肿瘤基因表达特征对肿瘤的诊断和治疗具有重要的生物学意义,而肿瘤与正常组织的分类是其中一个重要应用。根据多类别肿瘤基因表达谱,提出了一种自动特征选择方法。首先,结合非参数方法和filter思想,利用决策序列的随机性度量基因的权值并排序;然后,采用相关信息熵进行冗余性排除,自动地选择出具有高分辨能力、低冗余度的特征基因子集。实验结果表明,提出的方法能从多类别肿瘤基因表达谱数据中自动选出30个具有良好分类能力的特征基因,且具有较高的正确识别率。

关键词 肿瘤基因表达谱,特征选择,随机序列,相关信息熵

中图法分类号 TP391 **文献标识码** A

Auto-selection of Informative Gene for Multi-class Tumor Gene Expression Profiles

GAO Juan WANG Guo-yin HU Feng

(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract In microarray analysis, the selection of informative gene is an essential issue for tissue classification and successful treatment because of its ability to improve the accuracy and decrease computational complexity. The ability of successfully distinguishing tumor from normal tissues using gene expression data is an important aspect of this novel approach for cancer classification. In this paper, a non-parameter method for autonomous selection of informative gene was proposed for processing multi-class tumor gene expression profile, which contained 218 tumor samples spanning 14 common tumor types, as well as 90 normal tissue samples, to find a small subset of genes for distinguishing tumor from normal tissues. At First, the randomness of a decision sequence was defined to measure gene importance based on the non-parameter method and filter algorithm. Then correlation information entropy was used to eliminate redundant genes and selected informative feature genes. As a result, 30 informative genes are selected as markers for making distinctions between different tumor tissues and their normal counterparts. Simulation experiment results show that the selected genes are very efficient for distinguishing tumor from normal tissues. In the end, several methods for informative gene selection were also analyzed and compared to validate the feasibility and efficiency of the proposed method for dealing with tumor gene expression profiles.

Keywords Tumor gene expression, Feature selection, Random sequence, Correlation information entropy

1 引言

在肿瘤的诊断及预测过程中,由于肿瘤发展机制的复杂性,使得利用传统方法难以全面展开研究,而利用基因表达谱研究人员可以在分子水平上实现对肿瘤类型及亚型的准确识别,对肿瘤的诊断和治疗具有重要的生物学意义[1,2]。肿瘤数据通常具有小样本、超高维的特点,即临床样本少而包含的基因有成千上万,其中包含有用信息的基因只占很小一部分,大量冗余基因和噪声利用原始数据构建分类器来对新样本进行预测,这不但会花费大量时间,还会降低分类效果。文献[3]曾指出,在基因数据分析中,其特征基因的选择方法往往比分类器的选择更重要。因此,如何有效选择出肿瘤基因表达谱的分类特征基因是当前生物学研究的重点课题,研究人员也展开了大量的研究[4-8]。

从信息学角度出发寻找肿瘤相关基因、发现肿瘤基因表达特征是目前的主流方向,而从肿瘤与正常组织样本的基因表达谱数据出发,选取样本分类特征基因作为肿瘤的分子特征是研究肿瘤基因表达谱分类特征基因选择问题的一个重要应用[9]。目前,已有的基因特征选择方法主要包括:缠绕法(Wrapper)、过滤法(Filter)和嵌入式法(Embedded)[10]。其中,filter法独立于分类器,Wrapper法和Embedded法与分类器结合作用。过滤法基于基因数据本身的内在结构信息,根

到稿日期:2012-03-15 返修日期:2012-05-07 本文受国家自然科学基金(61073146),中国与波兰政府间科技合作项目(国科外字[2010]179号),重庆市教育委员会科学与技术研究项目(KJ110522)资助。

高 娟(1988-),女,硕士生,主要研究领域为智能信息处理、海量数据挖掘,E-mail:juan112087@163.com;王国胤(1970-),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为Rough集理论、粒计算、数据挖掘、知识技术等;胡 峰(1978-),男,博士,副教授,主要研究领域为智能信息处理。

据某个判别标准来对基因属性进行排序，不依赖于分类器对子集的评价，计算复杂度低，适合大规模的基因数据处理，并被证实有效。如 SNR[11]、统计量[12]、Relief[13]、Entropy[14]等 filter 型算法在特征选择领域得到了广泛应用。

但多数基于过滤法的基因特征选择算法都只考虑单个基因与相关类的关联度，忽略了基因之间的相似而造成冗余基因存在。Guyon[15]指出高度相关的特征是非常冗余的，因为我们不能从中获取对分类有用的信息。因此，肿瘤基因表达谱特征基因的选择既需要考虑基因的分辨能力，又需要考虑基因之间的冗余度。同时，非参数方法由于不需要假定一个具体的数据分布，因此比较适于用来分析基因表达数据[16]。基于以上分析，本文以多类别肿瘤基因表达谱作为具体的分析对象，从随机序列出发，提出通过度量决策序列的“随机性”来对基因属性进行排序，然后结合文献[17]提出的相关信息熵进行冗余去除，同时自动筛选出肿瘤特征基因。利用该方法，本文得到了 30 个具有较好分类性能的特征基因作为肿瘤表达数据的基因特征，然后采用支持向量机作为分类器进行分类测试，并与已存在的基因特征选择方法进行分析比较以说明该方法的有效性。

本文第 2 节对决策序列的随机性进行介绍；第 3 节结合序列随机性以及相关信息熵冗余性排除，给出了肿瘤基因特征选择算法；第 4 节在肿瘤基因表达谱数据集上进行了实验，并分析实验结果，得出结论。

2 基于序列随机性的基因排序方法

肿瘤基因表达谱每个样本都记录了组织细胞中所有可测基因的表达水平，从信息学角度讲，每个基因就是样本的一个属性，每个样本的类别信息就是该样本的决策属性。目前存在的过滤法，在类内变化小和类间差距大的思想下，依据某个判别标准来判断基因的重要性时，首先从决策属性出发，对决策属性进行排序，即按样本类别进行分组；然后以不同的方法来度量组内与组间间隔距离，从而衡量每个基因对类别的重要性。而“类内变化小，类间差距大”这一先验知识在高维空间普遍成立，但是对于低维空间，尤其是衡量单个基因对类别的分辨能力时不一定成立。例如，假设对于某个基因，其表达值过低或过高都会引发病变，只有表达值适中才不会引发病变，那么类内变化小和类间差距大策略便会带来较大的误差，不能准确评估基因重要性。基于以上分析，本文首先对条件属性（每个基因）进行排序，然后考虑对每个基因排序后，利用决策序列的“混乱程度”来衡量基因的重要性。本文采用序列的“随机性”来度量决策序列的“混乱程度”。

2.1 随机序列的概念

首先，假设一只口袋装有 12 只球，白球和红球各 6 只。从袋中随机将球取出，重复实验 3 次，得到下面 3 个序列（其中，0 代表白球，1 代表红球）：

Seq_1=〈000000111111〉；

Seq_2=〈000111111000〉；

Seq_3=〈011110011000〉。

我们可能主观认为 Seq_3 是随机的，Seq_1 不是随机的，即序列的“随机性程度”：$rand(Seq_1)<rand(Seq_2)<rand(Seq_3)$。但实际上各序列是等概率随机出现的[18]，各序列出现的概率均为 $6!*6!/12!$。下面给出随机序列的概念。

随机序列（Random Sequence）被定义成一组序列 $Seq=\langle x_1,x_2,\cdots,x_n\rangle$，其中：$x_i\in\{c_1,c_2,\cdots,c_m\}$，序列中类别为 c_k 的元素个数为 n_k，则所有不同的序列个数为 no. of $Seq=n!/n_1!\,n_2!\cdots n_k!$，每种序列等概率为 $prob(Seq)=1/\text{no. of } Seq$。

现在假设 Seq_1，Seq_2，Seq_3 分别是对基因 g_1、g_2、g_3 排序后得到的决策序列（0 代表病变，1 代表正常）。对于基因 g_1，其表达值过低会引发病变，过高不会引发病变；对于基因 g_2，其表达值过低或过高会引发病变，适中不会引发病变；对于基因 g_3，其表达值并不会对病变产生影响。而对于一个有分辨能力的基因，其基因表达值的不同取值范围总会表达不同的含义，对应着决策序列的不同取值，此时决策序列应具有规律性，而不是“随机”的，如 Seq_1 和 Seq_2。相反，对于一个没有分辨能力的基因，其基因表达值的不同取值范围对应的决策序列往往是无规律可言的，总是“随机”的，如 Seq_3。因此决策序列的这种“随机性”可以用来衡量基因的重要程度。

2.2 序列的随机性

定义 1（序列随机性 $rand(Seq)$） 随机序列 $Seq=\langle x_1,x_2,\cdots,x_n\rangle$，其中：$x_i\in\{c_1,c_2,\cdots,c_l\}$，$c_k$ 表示第 k 个类别，序列中类别为 c_k 的元素个数为 n_k。序列 Seq 对应的序列段为 $Seq=\langle seg_1,seg_2,\cdots,seg_r\rangle$，其中，序列段 seg_j 中元素的个数为 m_j，$m_1+m_2+\cdots+m_r=n$，$seg_j=\langle x_i,x_{i+1},\cdots,x_{i+m_j-1}\rangle$，且 $x_i=x_{i+1}=\cdots=x_{i+m_j-1}$，$1\leqslant j\leqslant r-1$，$1\leqslant i\leqslant n-m_j+1$，$class(seg_j)\neq class(seg_{j+1})$，$class(seg_j)$ 为 seg_j 对应的类别。序列 Seq 的随机性 $rand(Seq)$ 定义如下：

$$rand(Seq)=\sum_{i=1}^{k}H(c_i) \tag{1}$$

$$H(c_i)=-\sum_{class(seg_j)=c_i}\frac{|seg_j|}{n_i}\log_2\left(\frac{|seg_j|}{n_i}\right) \tag{2}$$

式(1)中，$H(c_i)$表示类别为 c_i 的样本在基因 x 的作用下的混乱程度。$H(c_i)$越小，类别为 c_i 的样本“随机性”越小，基因的分辨能力就越高；相反，$H(c_i)$越大，类别为 c_i 的样本“随机性”越大，基因就越不具有分辨能力。

例 1 利用定义 1 计算基因 g_1、g_2、g_3 的分辨能力。由前面可知，对基因 g_1、g_2、g_3 排序后得到的决策序列（0 代表病变，1 代表正常）分别为：

Seq_1=〈000000111111〉；

Seq_2=〈000111111000〉；

Seq_3=〈011110011000〉。

基因 g_1 对应的决策序列 Seq_1=〈000000111111〉对应 2 个序列段，即 $Seq_1=\langle seg_1,seg_2\rangle$，其中，$seg_1$=〈000000〉，$seg_2$=〈111111〉，则 Seq_1 的随机性为：

$$\begin{aligned}rand(Seq_1)&=\sum_{i=1}^{2}H(c_i)\\&=-\sum_{class(seg_j)=0}\frac{|seg_j|}{n_0}\log_2\left(\frac{|seg_j|}{n_0}\right)-\\&\quad\sum_{class(seg_j)=1}\frac{|seg_j|}{n_1}\log_2\left(\frac{|seg_j|}{n_1}\right)\\&=-\left(\frac{6}{6}\log_2\frac{6}{6}\right)-\left(\log_2\frac{6}{6}\right)\\&=0\end{aligned}$$

同理，Seq_2=〈000111111000〉的随机性为：

$$rand(Seq_2)=-(\frac{3}{6}\log_2\frac{3}{6}+\frac{3}{6}\log_2\frac{3}{6})-(\frac{6}{6}\log_2\frac{6}{6})$$
$$=1$$

Seq_3=〈011110011000〉的随机性为：

$$rand(Seq_3)=-(\frac{1}{6}\log_2\frac{1}{6}+\frac{2}{6}\log_2\frac{2}{6}+\frac{3}{6}\log_2\frac{3}{6})-$$
$$(\frac{4}{6}\log_2\frac{4}{6}+\frac{2}{6}\log_2\frac{2}{6})$$
$$=2.38$$

对上例中的基因 g_1、g_2、g_3，有 $rand(Seq_1)<rand(Seq_2)<rand(Seq_3)$，因此，基因重要性 $w(g_3)<w(g_2)<w(g_1)$，即基因 g_1 的分辨能力大于 g_2 和 g_3，基因 g_3 的分辨能力最小。显然，由定义 1 得到的序列的随机性能很好地反映出基因的鉴别能力，从而有助于获得理想的鉴别基因。

进一步地，我们再计算对基因 g_4 和 g_5 排序后，得到的决策序列的随机性，具体如表 1 所列。

表 1　不同基因对应的决策序列的随机性

随机序列	序列随机性
Seq_1=〈000000111111〉	rand(Seq_1)=0
Seq_2=〈000111111000〉	rand(Seq_2)=1
Seq_3=〈011110011000〉	rand(Seq_3)=2.38
Seq_4=〈001010110101〉	rand(Seq_4)=2.78
Seq_5=〈000111100011〉	rand(Seq_5)=1.92

从表 1 不难发现，对基因排序后得到的决策序列的随机性，能够很好地反映决策序列的混乱程度和该基因的分辨能力，以此判定基因的重要性具有可行性。

3　肿瘤基因表达谱特征基因选择方法

文献[19]指出，一个好的特征选择算法应该尽可能地合理高效，并能找到所含特征基因个数较少的典型基因组，也即是用尽可能少的基因来表达所包含的信息。因此在对肿瘤基因表达谱进行特征基因选择时必须有效去除冗余基因，使所选择的属性子集与决策类不仅具有较强的关联度，而且属性间的冗余度最小。为有效去除冗余基因，本文采用文献[17]提出的相关信息熵进行冗余性去除。

3.1　基于相关信息熵的冗余基因排除

我们知道随机变量的相关系数矩阵反映了变量相互间的线性相关程度，文献[20]通过分析指出 n 元随机变量 x_1，x_2，…，x_n 的线性相关性由均方误差 e 衡量：

$$e=a^{\mathrm{T}}Ra=y^{\mathrm{T}}\Lambda y=\lambda_1 y_1^2+\cdots+\lambda_n y_n^2\geqslant 0$$

当变量的线性组合为常系数方程时，e 的大小由特征值 λ_1，λ_2，…，λ_n 决定，特征值越小，则 e 越小，即一定程度上相关系数矩阵的特征值反映了变量的线性相关程度。文献[17]基于以上分析从信息论的角度出发提出了相关信息熵(Correlation Information Entropy)来进行冗余性排除，其定义如下：

$$H_R=-\sum_{i=1}^{N}\frac{\lambda_i}{N}\log_N\frac{\lambda_i}{N} \tag{3}$$

式中，λ_i 代表所选特征子集相关系数矩阵的第 i 个特征值，N 为特征子集中所含属性的数目。H_R 越大，即相关信息熵越大，则所选属性集的相关性越小，也即独立性越大；反之，亦然。如果所有属性线性相关，则相关熵为 0；如果所有属性均相互独立，则相关熵为 1。

为有效去除所选择特征基因子集的冗余性，应使所选择的属性集具有最小冗余度，即所选择的特征基因子集 S 具有最大的相关信息熵，亦即：

$$\max H_R(S\cup g_i),\forall g_i\in Gene-S \tag{4}$$

3.2　肿瘤基因表达谱特征选择算法

结合以上分析，本文针对肿瘤基因表达谱提出一种非参数的特征基因选择算法。算法思想如下：本算法首先利用序列的随机性来衡量每个基因的重要性，并排序；然后采用启发式前向搜索，特征基因子集 S 初始为空集，每次选择权值最大的基因添加到 S 中，如果该基因使子集的相关信息熵增大，即冗余性减少，则保留该属性，否则去除该属性，如此重复添加直至相关信息熵不再增大，算法停止。

算法具体描述如下。

算法 1　肿瘤基因表达谱的自动特征选择算法

输入：标准化后的肿瘤基因表达谱：

Gene=$\{g_1,g_2,\cdots,g_n\}$。

输出：特征基因集合 S。

```
Step1  // 初始化特征子集 S 和相关信息熵 H_R
       S=ϕ, H_R=0;
Step2  // 计算对基因排序后相应决策序列的随机性 rand(seq)_gene 并排序;
       for i=1 to n
         seq_i=sort(g_i); // 对每个基因进行排序
         rand(seq_i)=CalcRand(seq_i); // 计算对 g_i 排序后决策序列的随机性
       end for
       Gene'=sort(rand(seq_i));
Step3  // 采用相关信息熵进行冗余性排除:
       for i=1 to n
         Calculate H_R(S∪g_i),(g_i∈Gene'-S);
                  // 将权值最大的基因添加到特征子集中,并计算其相关信息熵
         if(H_R(S∪g_i)-H_R(S))>0
         then S=S∪g_i
         end if
         //如果相关信息熵增大,就将该基因添加到特征基因集合中,否则去掉
       End for
Step4  Return S
```

现分析算法的复杂性，假设总共 n 个基因，容易计算整个算法的时间复杂度：计算决策序列的随机性只需要对序列扫描一次，统计各序列段大小，再按权值大小对基因排序，其时间复杂度为 $O(n\log n)$；再利用相关信息熵进行冗余性排除，其时间复杂度为 $O(n)$，因此，总的时间复杂度为 $O(n\log n)$。

4　实验及结果分析

4.1　肿瘤基因表达谱数据描述

本文采用美国麻省理工学院提供的多类别肿瘤基因表达谱数据集[21]（下载网址：http://www.broad.mit.edu/cgi/cancer/datasets.cgi），该数据集共包含 308 个样本，其中 218 个样本为肿瘤组织样本，涵盖了目前常见的 14 种不同组织类型的肿瘤；90 个样本为对应组织的正常样本。每个样本都记录了组织细胞中 16063 个基因的表达水平。

针对该数据集，首先对每个基因表达水平进行标准化(均值为0，方差为1)，然后采用文献[9]的测试方法，将不同肿瘤组织作为一个“整体”，考虑其与正常组织在基因表达水平上的区别，把整个数据集划分为训练集 S_{train} 和测试集 S_{test} 两部分，如图1所示，每种类型的肿瘤与其对应的正常组织均按近似于2∶1的比例分配在 S_{train} 和 S_{test} 中。文献[9]指出由于在划分数据样本时，考虑了训练集和测试集都包含所有不同类型的肿瘤组织及其对应的正常组织样本，因此在训练集上选择得到的特征基因能反映出整个“肿瘤基因”与“正常组织”在基因表达上的差异。

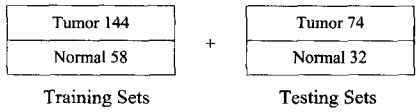

图1 数据样本集的划分

4.2 分类器

SVM (Support Vector Machine，支持向量机)主要针对两类线性可分问题，其基本原理是寻找一个最优分类面，并使其两侧的分类间隙最大，即在特征空间中构建最优分割超平面 $w^Tx+b=0$ 使得：$w^Tx+b\geqslant 1, y=1$；$w^Tx+b\leqslant -1, y=-1$；将其合并为 $y(w^Tx+b)\geqslant 1$，并最大化超平面与不同类别样本集之间的距离，即最小化 $w^Tw/2$。本文针对肿瘤基因表达谱数据集，分类器采用 LIBSVM[21]，核函数 $K(x_i,x_j)=x_i\cdot x_j$，即将SVM用作线性分类器，惩罚因子 $C=100$，其他参数默认。

4.3 实验分析

本文针对肿瘤基因表达谱数据，提出的特征选择方法主要分为两步：第一步根据 filter 型算法思想，利用决策序列的随机性计算每个基因的权值，并按权值大小排序；第二步，从基因集合出发，利用相关信息熵去除冗余，并自动确定特征基因的数目。通过这两个步骤，我们筛选出了最有分辨能力，且冗余度最低的30个特征基因 F_s，这30个基因对测试集 S_{test} 中106个样本的总的错分数为3.78，分类正确率达96.44%。

为了检验本文所提算法对肿瘤基因表达谱数据处理的有效性，从以下4个方面对选择得到的30个特征基因的分类能力进行检验。

4.3.1 特征基因的分辨能力

基因特征选择的一个重要目的是检验基因是否具有分辨能力，即检验这些基因在肿瘤与正常组织样本中的表达水平是否显著不同。本文选择得到的30个特征基因在肿瘤和正常组织样本中的平均表达水平如图2所示。

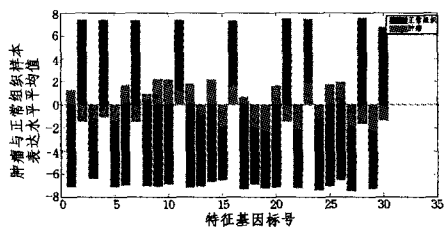

图2 特征基因在肿瘤与正常组织样本中的差异表达

从图2中可以看出这些特征基因在肿瘤与正常组织样本中的平均表达水平有着显著不同，说明了经本文算法选择得到的这些特征基因具有较强的分辨能力；同时也说明了尽管不同的肿瘤组织在细胞形态和组织结构表现各异，然而从分子水平上看，不同肿瘤在其基因表达水平值上确实存在着较大程度的共性，从而证明了在分子水平上实现对肿瘤的诊断和治疗具有可行性。这有利于区分正常组织与肿瘤组织，为临床治疗做出早期诊断，为进一步研究理解肿瘤的发生机制以及肿瘤的临床治疗提供了重要依据。

4.3.2 特征基因子集的冗余性分析

基因特征选择的另一个重要目标就是在不降低分类器分辨能力的前提下，选择最少的、最有“鉴别”能力的，且能够代表整个基因全集的特征子集。冗余基因的存在并不能使我们获得更多有用的信息，反而会增加计算的复杂度和导致分类器性能降低。本文选择得到的30个特征基因子集的相关信息熵如图3所示。

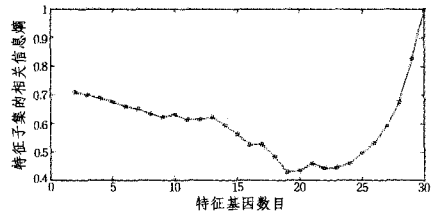

图3 特征基因子集的冗余性

由图3知，利用本文算法得到的30个特征基因子集的相关信息熵接近于1。根据式(3)可以知道，此时特征基因子集的相关性很小，特征基因具有高度的独立性，因此本文算法有效去除了冗余基因，得到了数量最少且分类能力最强的分类特征基因。

4.3.3 分类性能

本文是按照文献[9]的方法对肿瘤基因表达谱数据进行训练集和测试集的划分，而该划分方法是一种随机划分，故将每种类型的肿瘤组织及其对应的正常组织均按近似2∶1的比例分配在训练集和测试集中有多种划分方法。为了检验本文算法的有效性和稳定性，避免特征基因集合 F_s 出现“过学习”现象，本文仍按文献[9]的检验方法进行检验，即基于样本抽样，利用随机测试实验的方法对这30个基因的分类能力进行检验。具体做法如下：在保持训练集和测试集大小不变、不同类型的肿瘤与正常组织在训练集及测试集中均按近似2∶1分布的条件下，从总体样本中采用无重复抽样的方式随机抽取样本形成新的训练集 S_{train}，剩余样本作为测试集 S_{test}；以前面得到的30个特征基因作为样本的分类特征，利用 S_{train} 训练SVM分类模型，对测试集 S_{test} 进行样本识别，并重复这样的操作500次，记录每次的测试结果，统计得到总的错分率均值为0.105，标准差为0.0336。同时，为了说明该算法的有效性，本文也将未经特征基因选择的肿瘤基因表达谱按上述步骤实验500次，记录每次的测试结果，将其作为对比，得到总的错分率均值为0.243，标准差为0.04276，具体如表2所列。

表2 自动特征选择分类对比结果

实验次数	分类器	未经特征选择的肿瘤基因表达谱数据		本文算法	
		总的错分数均值	总的正确率均值	总的错分数均值	总的正确率均值
500	SVM	25.7	75.7%	11.1	89.5%

从表2可以看出，本文算法选择得到的特征基因集合对肿瘤组织和正常组织确实有很好的分类能力，从而比未经过

特征选择的基因分类准确率有很大提高，说明本文算法得到的30个特征基因可以代表肿瘤基因表达谱，能够保持整个基因数据集的分类能力，从而证明本算法可行有效。

4.3.4 不同基因选择方法结果的差异性

为进一步验证本文算法的有效性和不同基因选择方法结果间的差异性，本文选择了Golub等人[11]提出的"信噪比"(Signal to Noise ratio，S2N)指标，即使用均值和方差构造的统计量作为鉴别基因的选择量度，该方法是目前确定特征基因的一个经典方法。另外，文献[12]使用两样本t统计量(two sample t-statistic)作为基因排序的标准进行基因选择，其广泛应用于基因表达谱的分析中。这些方法是经典的filter型算法，但较明显的不足是没有去除特征子集的冗余性。Relief算法[13]在一定程度上考虑了属性间的相互关系，因此，本文利用信噪比方法、t-test、Relief和文献[9]提出的RFE-Relief方法对肿瘤基因表达谱数据进行特征选择，仍采用SVM作为分类器，分类结果如表3所列。

表3 5种分类方法的分类性能比较

Method	特征基因数目	最小分类错误次数	时间复杂度
信噪比	1957	11	O(n)
T-test	1158	12	O(n)
Relief-A	367	11	O(n)
RFE-Relief	98	3	$O(n^2)$
本文算法	30	4	O(nlogn)

文献[19]指出，进行特征选择的目标是用最少的基因得到最大的分类正确率，同时还得兼顾较小的时空开销。从表3可以看出，RFE-Relief方法的最小分类错误率最小，但该方法需要很大的计算量，其计算复杂度为$O(n^2)$，而肿瘤基因表达谱数据维数过高，因此，该方法在寻找特征子集时空间开销过大。而本文算法的最小分类错误率大大高于信噪比、t-test、Relief等算法，并且能自动确定特征基因的数目，同时与其它4种方法相比，本文算法选择得到的特征集合所含特征基因最小，但其分类能力强，且计算复杂度比RFE-Relief算法小，因此，本文针对肿瘤基因表达谱提出的结合决策序列的随机性和相关信息熵特征选择算法效率高，有效性强。

结束语 基因特征选择是分析基因表达谱的一个核心内容，它既是建立有效分类模型的关键，也是发现疾病基因的重要手段，可以去掉无用的噪声基因，提高基因分类器的识别率，降低计算复杂度。

本文结合单个基因的分类能力和基因之间的相关性，对肿瘤基因表达谱特征基因选择问题进行了研究。首先从决策序列的随机性出发，计算每个基因的权值，并依据权值进行排序；然后采用相关信息熵进行冗余性排除，该方法是一种非参数方法，不需要假定肿瘤基因表达谱数据服从某一具体分布，并能自主选择特征基因，确定特征基因的数目。实验证明，本文得到的特征基因集合具有较大的分类能力，说明了本文所提策略的有效性。以此希望帮助生物医学研究者选择有用的基因，进而指导完成癌症等病例的诊断、研究与预测。

参考文献

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

[2] Ramaswamy S, Golub T R. DNA microarrays in clinical oncology[J]. Jornal of Clinical Oncology, 2002, 20(7): 1932-1941

[3] Krishnapuram B, Carin L, Hartemink A. Gene expression analysis: Joint feature selection and classifier design[M]// Scholkopf B, Tsuda K, Vert J-P, eds. Kernel Methods in Computational Biology, MIT, 2004: 299-318

[4] 李建更，高志坤. 随机森林：一种重要的肿瘤特征基因选择法[J]. 生物物理学报，2009，25(1)：51-56

[5] 徐菲菲，苗夺谦，魏莱. 基于模糊粗糙集的肿瘤分类特征基因选取[J]. 计算机科学，2009，36(3)：196-200

[6] 李颖新，阮晓钢. 基于支持向量机的肿瘤分类特征基因选取[J]. 计算机研究与发展，2005，42(10)：1796-1801

[7] 李泽，包雷，黄英武，等. 基于基因表达谱的肿瘤分型和特征基因选取[J]. 生物物理学报，2002，18(4)：413-417

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

[9] 李颖新，李建更，阮晓钢. 肿瘤基因表达谱分类特征基因选取问题及分析方法研究[J]. 计算机学报，2006，29(2)：324-330

[10] Saeys Y, Inza I, Larran-aga P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23 (19): 2507-2517

[11] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression[J]. Science, 1999, 286: 531-537

[12] Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer[J]. New England Journal of Medicine, 2001, 344(8): 529-548

[13] Kononenko I. Estimating Attributes Analysis and Extensions of RELIEF[J]. Machine Learning: ECML-94 Lecture Notes in Computer Science, 1994, 784: 171-182

[14] Zhu S H, Wang D, Yu K, et al. Feature Selection for Gene Expression Using Model-Based Entropy[J]. IEEE Transaction on Computational Biology and Bioinformatics, 2010, 7(1): 25-36

[15] Guyon I, Weston J, Barnhill S, et al. Gene Selection for Cancer Classification using Support Vector Machines [J]. Machine Learning, 2002, 46(13): 389-422

[16] 李建中，杨昆，高宏，等. 考虑样本不平衡的模型无关的基因选择方法[J]. 软件学报，2006，17(7)：1485-1493

[17] Qiang W, Yi S, Ye Z, et al. A Quantitative Method for Evaluating the Performances of Hyperspectral Image Fusion[J]. IEEE Transactions on Mentation and Measurement, 2003, 52 (4): 1041-1046

[18] Volchan S B. What is a random sequence[J]. The American Mathematical Monthly, 2002, 109: 46-63

[19] 周昉，何洁月. 生物信息学中基因芯片的特征选择技术综述[J]. 计算机科学，2007，34(12)：143-150

[20] 章舜仲，王树梅. 相关系数矩阵与多元线性相关分析[J]. 大学数学，2011，27(1)：195-198

[21] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(26): 15149-15154

[22] Chang C-C, Lin C-J. LIBSVM [EB/OL]. http://www.csie.ntu.edutw/~cjlin/libsvm