doi:10.3969/j.issn.1672-5565.2013.01.12

基于SVM和平均影响值的人肿瘤信息基因提取

李凌波,张 静*,陈 丹
(云南大学数学与统计学院,昆明 650091)

摘 要:基于基因表达谱的肿瘤分类信息基因选取是发现肿瘤特异表达基因、探索肿瘤基因表达模式的重要手段。借助由基因表达谱获得的分类信息进行肿瘤诊断是当今生物信息学领域中的一个重要研究方向,有望成为临床医学上一种快速而有效的肿瘤分子诊断方法。鉴于肿瘤基因表达谱样本数据维数高、样本量小以及噪音大等特点,提出一种结合支持向量机应用平均影响值来寻找肿瘤信息基因的算法,其优点是能够搜索到基因数量尽可能少而分类能力尽可能强的多个信息基因子集。采用二分类肿瘤数据集验证算法的可行性和有效性,对于结肠癌样本集,只需3个基因就能获得100%的留一法交叉验证识别准确率。为避免样本集的不同划分对分类性能的影响,进一步采用全折交叉验证方法来评估各信息基因子集的分类性能,优选出更可靠的信息基因子集。与基它肿瘤分类方法相比,实验结果在信息基因数量以及分类性能方面具有明显的优势。

关键词:基因表达谱,秩和检验,支持向量机,平均影响值,全折交叉验证
中图分类号:Q811.4 **文献标识码**:C **文章编号**:1672-5565(2013)-01-072-07

Selection of human tumor information genes based on the support vector machine and mean impact value

LI Ling-Bo, ZHANG Jing*, CHEN Dan
(*School of Mathematics and Statistics*, *Yunnan University*, *Kunming* 650091, *China*)

Abstract: Selection of information genes for tumor classification based on gene expression profiles is a main means to find specific expression genes and to study their expression pattern. Tumor diagnosis via the information genes obtained from gene expression spectrum is becoming an important research field of bioinformatics and is expected to be a fast and effective method for molecular diagnosis of tumors in clinical medicine. Considering the characteristics of gene expression profiling data of tumors such as high dimensions, small sample size and large noise etc, an algorithm for searching information genes is proposed that exploits support vector machine (SVM) and combines mean impact value (MIV). The advantage of this algorithm is that more information gene subsets with less genes and powerful classification capacity could be searched. A binary classification tumor dataset is applied to examine this novel algorithm, the result shows that it is feasible and effective in tumor classification. For colon cancer sample set, only 3 genes can reach 100% accuracy of leave-one-out cross validation (LOOCV). To avoid the influence of classification performance because of the different partition for the sample set, full cross validation method is further used to assess the classification performance of the information gene subsets. More credible information gene subsets are selected. Compared with other tumor classification methods, the result is superior both in information gene number and in classification capacity.

Key words: Gene Expression Profile, Rank-sum Test, Support Vector Machine, Mean Impact Value, Full-fold Cross Validated

---

收稿日期:2012-11-10;**修回日期**:2012-11-27.
基金项目:国家自然科学基金项目(11261066),云南省应用基础研究资助项目(2007A023M),云南省教育厅科学研究项目(2012Y497)。
作者简介:李凌波,女,湖南邵阳市人,硕士,研究方向:生物信息学,E-mail:361990236@qq.com.
*通讯作者:张静,Tel: 13700654882,E-mail: zhangjing@ynu.edu.cn.

DNA 芯片技术的发展产生了大量与人类疾病基因相关的表达谱数据,其中肿瘤基因表达谱受到广泛关注。由于肿瘤基因表达谱数据存在样本量少、维数过高以及噪声大等特点,使其在统计分析及生物医学处理上都遇到一些困难。因此在肿瘤基因分类前必须进行特征选择以降低维数,从而便于发现对样本类别有决定或重要影响的基因。这些基因对肿瘤疾病研究具有重要意义:一方面特征基因的有效选取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,同时也为肿瘤药物治疗分子靶标的确定从生物信息学角度提供线索;另一方面,对这些特征基因表达行为的分析将有助于揭示肿瘤发生与发展的分子机制,使人们将注意力从整个基因组范围集中到有限少数几个基因上,从而使实验研究更具有针对性。

特征选择是高维数、高噪声、小样本数据面临的主要问题,目前常用的特征选择方法为过滤(Filter)法有监督学习的缠绕(Wrapper)法,前者直接从数据自身的特点出发,运行效率较高,但完全独立于最终预测的分类器,不利于优化分类性能;有监督缠绕法能获得较高的分类率,但 Wrapper 是一个循环反馈改进的过程,运行效率较低,尤其当数据具有大噪声和"维数灾难"时,容易引起监督学习模型的过拟合。结合二者以取长补短的 Filter – Wrapper 混合方法逐渐受到关注,该方法首先采用 Filter 方法从大量的变量中初选出一定数量的备选变量以大幅降低特征变量的搜索空间,然后再用 Wrapper 法,以分类精度为指标精选出满足目标条件的特征变量。

支持向量机(Support Vector Machine,SVM)是一种小样本学习理论,适合处理基因表达谱这种样本量少、变量维数高的数据集分类和特征选取问题。支持向量机是由 Vapnik[1] 等人基于统计学习理论,采用结构风险最小化原理提出的一种机器学习算法,通过调整判别函数使得它最好地利用边界样本点的分类信息,构造出最佳分类超平面。Brown[2] 等将支持向量机、神经网络、近邻等几种常用分类方法分别应用到基于基因表达谱的肿瘤基因分类,并对分类效果进行比较发现采用支持向量机作为分类器效果最好。本文在分析肿瘤基因表达谱样本集的特点的基础上提出一种基于 SVM 的 Filter – Wrapper 混合特征选择方法,其实质是以 SVM 分类性能为评估准则的寻找特征基因的启发式宽度优先搜索算法。通过对结肠癌数据集进行实验分析,证明该算法是可行且有效的。

1　材料与方法

1.1　材料

分析样本采用 Alon 等[3] 公布的结肠癌基因表达数据集。该数据集属于二分类肿瘤数据集,包括 40 个肿瘤组织和 22 个正常组织样本,每个样本均含 2000 个基因表达数据。

1.2　方法

1.2.1　基于非参数检验的肿瘤信息基因初选

由于基因数目较大,而大多数基因的表达又与肿瘤无关,因此在判断肿瘤基因标签的过程中先要剔除大量的"无关基因"以缩小搜索致癌基因的范围。为此先引入一个基因排序辨识性度量对原始基因集合进行粗选,选出对肿瘤辨识性较大的基因。为了避免正态性假设(邓林等[4] 对结肠癌、白血病和乳腺癌数据集的统计分析表明这三种肿瘤数据集都不服从正态分布),本文采用 Wilcoxon 秩和检验对两总体微阵列数据结肠癌数据集进行信息基因初选。Wilcoxon 秩和检验[5] 是一种建立在二项分布理论基础上的总体分布位置差异非参数检验方法,它根据基因表达数据的大小排序,然后得到数据的秩,再利用数据的秩而不是数据本身计算基因的秩和统计量,并根据统计量筛选出在各总体中表达差异最显著的一些基因,供后续肿瘤分类模型的建立及分类特征选取使用。

1.2.2　基于 SVM 的基因平均影响值分析

基于基因排序的信息基因初选结果一般还不能达到信息基因子集期望维数的要求(一般只需 3 或 4 个信息基因,分类的效果会更好[6]),需要进一步采用 Wrapper 方法[10],以分类精度为指标精选出满足目标条件的信息基因子集。基于 SVM 的基因平均影响值(Mean Impact Value,MIV)可用来分析各基因影响 SVM 模型输出能力的大小,MIV 被认为是神经网络中评价变量相关性最好的指标之一[11],本文以此为基础设计一种结合 SVM 应用平均影响值的方法(SVMMIV)来精选信息基因。

MIV 作为评价各个自变量对因变量影响的重要性指标,其符号代表相关的方向,绝对值大小代表影响的重要程度。具体做法为:在训练 SVM 模型终止后,将训练样本 P 中每一个自变量特征在其原值的基础上分别加、减构成新的两个训练样本 P_1 和 P_2,将 P_1、P_2 分别作为测试集利用已建成的 SVM 模型进行仿真,得到两个仿真结果 A_1 和 A_2,A_1 和 A_2 的差值即为变动该自变量后对输出产生的影响变化值(IV,Impact Value),将 IV 按观测例数平均得出该自

变量对于因变量的 MIV。依次算出各个自变量的 MIV 值,根据 MIV 绝对值的大小为各自变量排序,依次去除对决策影响最小的若干个基因,将剩下的基因视为候选信息基因子集,考察它对样本的分类能力,从中找出具有最佳分类能力且所含基因最少的候选信息基因子集作为分类信息基因集合。每去除一次基因,都要重新训练 SVM 模型,获取新的决策函数,并计算剩下基因对决策的 MIV 值,具体步骤为:

(1) 在训练集中用候选信息基因子集 P 训练 SVM 模型,并记录 P 在"留一法"交叉验证和"独立测试实验"时分类准确率;

(2) 计算 P 中各基因对决策的 MIV 值;

(3) 找出 P 中 MIV 值最小的若干基因 g,再从 P 中去除这些基因,得到新的候选信息基因子集:$P = P - \{g\}$;

(4) 若($size(P)$为 matlab 软件中返回矩阵 P 的列数),则返回步骤(1)继续执行,否则退出。

1.2.3 支持向量机

支持向量机(*SVM*)用来区分具有标记的两类样本,其构建分为两个部分:一部分为训练部分,另一部分为测试部分。对给定的训练样本集合$\{(x_1, y_1\},(x_2,y_2),\cdots,(x_1,y_1)\}$其中 $x_i \in R^n, i = 1,\cdots,l, x$ 是训练样本中的元素,其维数为 n;$y_i \in \{-1,1\}, i = 1,\cdots,l, y_i$ 是样本 x_i 归属的类别。如果 x_i 是负类元素,则 $y_i = -1$;若 x_i 是正类元素,则 $y_i = 1$。在肿瘤分类领域常用 *Gauss* 径向基核函数(*Radial Basis Function*,*RBF*)[12],此时最佳超平面满足约束条件:

$$max\ L(\alpha) = \sum_{i=1}^{m}\alpha_i - \sum_{i,j=1}^{m}\alpha_i y_i K(x_i,x_j)$$

$$K(x,y) = \exp(-\| x-y \|^2)/(2\delta))$$

$$\sum_{i=1}^{m}\alpha_i y_i = 0$$

其中 $\alpha_i \in [0,C]\ (C>0)$ 是拉格朗日乘子,$i = 1,2,\cdots,m, K(x_i,x_j)$ 是核函数,b 是一个偏置量。对于测试样本 x,根据判别函数 $f(x) = \mathrm{sgn}(\sum_{i=1}^{m}\alpha_i y_i K(x_i,x) + b)$ 的正负来决定样本的类别。本文的实验将采用径向基函数分类器(*SVM*(*RBF*))进行肿瘤分类特征选择和分类精度检验,*SVM* 的实现使用的是 *LIBSVM* 工具箱[13]。

1.2.4 肿瘤分类模型的评估

在肿瘤分类领域,通常采用 k - 折交叉验证方法(k - fold Cross Validation,k - fold CV)来评估分类模型的泛化性能,然而值的选取往往会影响分类的准确率。由于肿瘤样本规模小,一些文献采用留一法来评估肿瘤分类模型,Breiman 等[14]认为折或折交叉验证方法优于留一法。王树林等[6]通过实验发现样本集的不同划分对分类准确率有一定的影响,提出一种能够客观反映分类器泛化性能的全折交叉验证评估方法:记 $CV(k)$ 表示样本集的折交叉验证分类准确率,其中 $2 \leqslant k \leqslant m$,$m$ 为样本总数,定义分类准确率均值为:

$$mean = \frac{1}{m-1}[\sum_{k=2}^{m} CV(k)]$$ 标准差为

$$std = \sqrt{\frac{\sum_{k=2}^{m}(CV(k) - mean)^2}{m-2}}$$

此时具有最大分类准确率均值与最小标准差的基因子集的泛化误差最小。用这种方法获得的分类准确率均值称为全折交叉验证分类准确率,标准差表示不同划分对分类准确率的影响程度。

在 SVMMIV 算法中直接采用全折交叉验证方法评估肿瘤分类模型会增加计算量,因此本文先用留一法(LOO,leave - one - out)评估分类器,搜索出所有具有最高 LOO 分类准确率的信息基因子集后,再采用全折交叉验证方法对其进行评估。

2 结 果

2.1 秩和方法的实验结果

为了检验秩和方法在特征基因选取方面的有效性,我们以选取的相关基因分类能力为依据,在结肠癌数据上利用秩和方法结合 SVM 进行实验。在一些常用的显著性水平下,采用秩和方法对数据集做了相关基因选取,亦即计算出每个基因的统计量以及相应的 p 值,与事先设定的某个常见的显著性水平 α 比较,若 p 值小于显著性水平 α 就将该基因选取出来,结果见表 1。

表 1 不同显著性水平下相关基因数目及分类准确率

Table 1 The number of related - gene and accuracy under different significance levels

显著性水平 α	0.1	0.05	0.01	0.001	0.0001	原数据
相关基因数	522	387	188	61	22	2000
分类准确率(%)	95	97.5	97.5	97.5	95	72.5

在选取的相关基因集合上,使用径向基 SVM 进行学习和预测。实验中,先对样本数据进行归一化,使每个基因的表达值均值为 0,方差为 1,然后将正常(Normal)和肿瘤(Cancer)样本按接近的比例随机地分配到训练集和测试集中,正常的分别为 14 和 8,而肿瘤的为 26 和 14。为了使实验结果更可靠,我们采用留一法进行检验,得到平均分类正确率,结

果见表 1。可以看出,利用秩和方法进行基因选取是有效的,在基因选取后得到的最好分类正确率是 97.5%,利用秩和方法进行相关基因选取能使肿瘤预测的准确率大幅度提高。

同时我们看到,显著性水平的选取是后续获得较高分类准确率的关键。显著性水平取得过高,一些与肿瘤因素相关性不显著的基因会被选取出来,成为干扰分类器的噪声,降低分类准确率;而显著性水平取得过低,又会漏掉一些有用的相关基因。从表 1 看,取 0.05、0.01、0.001 的显著性水平都能够使结肠数据获得最好的分类结果,但是在 0.05 显著性水平下所取得的相关基因数最多,因此包含的信息量也最大。

2.2　基于 SVM 的基因平均影响值分析结果

在 0.05 的显著性水平下,选取出的相关基因有 387 个。通过实验发现训练集和测试集的不同划分对得到的 SVM 模型具有一定的影响,因此在采用 SVMMIV 算法进行信息基因精选时,其结果可能会因训练集和测试集的不同划分而变动。为了消减这种影响,本文在采用 SVMMIV 算法进行信息基因精选时进行了 10 次独立重复实验,每次实验前均将正常和肿瘤样本数目按接近的比例随机地分配到训练集和测试集。设定阈值为 90%,分别选出分类准确率大于 90% 的两基因子集、三基因子集与四基因子集。但由于所有实验结果都是使用“留一法”交叉验证方法获得的,其分类准确率可能受样本集的不同划分影响,所以进一步采用全折交叉验证方法评估各信息基因子集的分类性能以优选出更客观的信息基因子集。

鉴于样本所属类别的标号,即 $y=1$ 表示为正常样本;$y=-1$ 表示为肿瘤样本 j,因此,对于样本,在 SVMMIV 算法中得到的两个仿真值 A_{1j} 和 A_{2j} 要么都为 1,要么都为,要么一个为 1 一个为,则 A_{1j} 和 A_{2j} 的差值,即各自变量变动对该样本输出产生的影响变化值要么是 0,要么是 ±2。显然在接下来求各自变量的 MIV 值时,会因为 ±2 相加为 0 而导致计算得到的 MIV 值小于各自变量对因变量的实际平均影响值。这样就会出现对输出影响程度较大的自变量的 MIV 值小于对输出影响程度偏小的自变量的 MIV 值的现象。为此,在求各自变量的 MIV 值时,将原来求 IV 值的简单算术平均数转化为求 IV 绝对值的简单算术平均数。表 2 列出了一次实验所得到的不同维数的特征基因子集及其分类准确率,从中可以看出,当特征基因子集的维数 n 小于 13 时,随着特征基因个数的减少,其分类性能反而增高,尤其当维数为 3 时,就可以获得最高为 97.5% 的分类正确率。这也符合王树林等的结论,即当特征较少(3 个或 4 个信息基因)时,分类的效果更好[6]。因此,基于 MIV 方法的支持向量网络对变量进行筛选是可行的。

表 2　各维特征基因子集的样本分类能力

Table 2　Classification performance of different subsets

特征基因子集维数	387	40	13	6	5	3	2
分类正确率(%)	97.5	95	87.5	92.5	95	97.5	95

MIV 作为评价变量对结果影响大小的指标,其大小代表影响的相对重要性。表 3 给出了一次实验计算得到的低维特征基因子集中各个基因的 MIV 值。当特征基因子集的维数为 6 时,该子集的分类正确率为 92.5%,其中第 1 ~ 5 个基因得出的 MIV 值较大,第 6 个基因的 MIV 值最小。剔除第 6 个基因,得到新的维数为 5 的特征基因子集,其分类正确率提高了 2.5%。再次计算各基因的 MIV 值,剔除 MIV 值最小的基因。由于该基因子集中第 2 个基因与第 5 个基因的 MIV 值相等且最小,因此将它们同时剔除。得到维数为 3,分类正确率为 97.5% 的特征基因子集,达到了既减少自变量又提高分类正确率的目的。此外可以看出当维数为 2 时,分类正确率又回落到 95%,由此可见,维数为 3 的基因子集为{J02854,M63391,M76378}即是该次实验寻求的维数最低、分类能力最高的特征基因子集,其全折交叉验证结果可参考表 4 第 5 组基因。

表 3　部分基因的 MIV 值

Table 3　Mean impact value of partial genes

n = 6			n = 5			n = 3			n = 2		
gene	MIV	rank	gene	MIV	rank	gene	MIV	rank	gene	MIV	rank
J02854	0.075	3	J02854	0.075	3	J02854	0.075	3	M63391	0.325	1
H43887	0.05	5	H43887	0.05	4	M63391	0.175	1	M76378	0.2	2
M63391	0.15	1	M63391	0.125	1	M76378	0.1	2	–	–	–
M76378	0.1	2	M76378	0.1	2	–	–	–	–	–	–
R87126	0.075	3	R87126	0.05	4	–	–	–	–	–	–
M82919	0.025	6	–	–	–	–	–	–	–	–	–

表 4 给出了用 Wilcoxon + SVMMIV 方法在肿瘤数据集上进行实验所获得的部分实验结果。三基因子集{R87126, K03460, R08021}虽然可获得 100% 的留一交叉验证分类准确率,但其全折交叉验证分类准确率为,表明采用该基因子集来分类样本,两类样本的边界比较模糊,样本集的不同划分对分类准确率有一定影响,其影响程度可通过相应的标准差来反映。但即便如此,通过比较不难发现三基因子集{R87126, K03460, R08021}的全折交叉验证分类准确率比其他的基因子集的全折交叉验证分类准确率都要高,表明这个基因子集在分类性能方面优于其它获得 100% 的留一交叉验证分类准确率的基因子集。图 1(A)显示了三基因子集{R87126, K03460, R08021}的各折交叉验证分类准确率,从图中可以看出除了 2,3,5,7,9 折交叉验证分类准确率低于 100%,其它各折交叉验证分类准确率均为 100%,这表明样本集的不同划分对分类准确率有一定影响。

表 4 本文方法的部分实验结果

Table 4 Partial results on Colon dataset using our method

序号	基因组合				留一法			CV 准确率%		
					C	g	best	mean	std	max
1	R87126	K03460	R08021		0.71	4	100	99.55	1.39	100
2	H23544	D42047			16	11.31	100	99.36	2.35	100
3	M26383	T47377	R08021		2	16	100	99.29	2.36	100
4	H24030	X12496	R08021	T51858	1.41	16	100	99.29	1.81	100
5	J02854	M63391	M76378		0.25	2.83	97.5	97.5	0	97.5
6	J02854	M63391	R87126		0.25	4	97.5	97.5	0	97.5
7	H06524	M58050	R62549	H24030	1.41	16	97.5	97.31	0.86	97.5
8	J02854	R08021	T70062		2.83	11.31	97.5	97.25	1.43	100
9	T60155	U09587	R08021		1.41	4	97.5	97.24	1.12	97.5
10	M82919	H43887	M80815	M76378	1.41	16	97.5	96.67	1.32	97.5
11	J02854	R08021	U30825		0.71	5.66	97.5	95.32	1.83	100

C and g 表示相应基因子集获得最高“留一法”交叉验证分类准确率时的参数取值,表示相应基因子集获得的最高“留一法”交叉验证分类准确率,表示相应基因子集获得的最高折交叉验证分类准确率。

图 1(B)显示了由三基因子集{R87126, K03460, R08021}表达数据构成的三维散点图,这种分类结果的可视化对于肿瘤临床诊断是很有意义的,它能帮助医务人员分析临床样本并积累诊断经验。基因的全折交叉验证分类准确率为 99.55%。

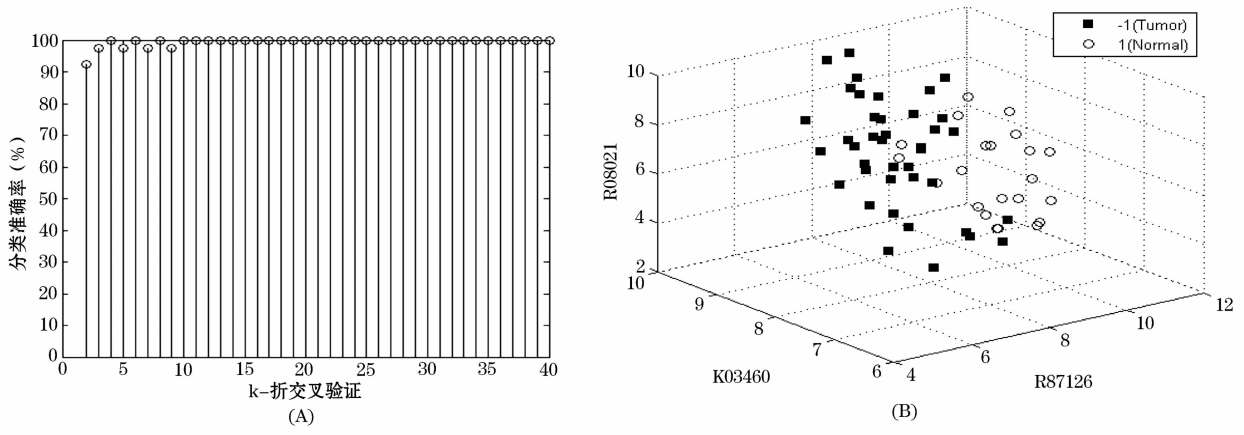

图 1 三基因子集{R87126, K03460, R08021}的各折交叉验证分类准确率
(mean = 99.55%, std = 1.39%)(A)及表达数据的三维散点图(B)

Fig. 1 The classified accuracy of genes{R87126, K03460, R08021} under each fold Cross Validation(A), and the three - dimensional scatter diagram constructed by the genes expression data(B)

表 5　针对结肠癌数据集的不同分类方法所获得的分类结果比较

Table 5　Comparison of Colon dataset using different methods

序号	实验方法		息基因数量	识别精度	参考文献
	基因选择或特征抽取方法	分类器			
1	秩和检验方法 SVMMIV 算法	SVM(RBF)	3	100%	本文
2	特征记分准则 (Feature Score Criterion,FSC)	SVM	–	90.30%	[15]
3	递归特征删除方法 (Recursive Feature Elimination,RFE)	SVM	4	98.0%	[16]
4	遗传算法 (Genetic Algorithms)	SVM	12	93.55%	[17]
5	多目标演化算法 (Multi – objective Evolutionary Algorithms)	基因表达 差异判别	7	97.0%	[18]
6	秩和基因选取方法 (Rank – sum test)	SVM	34	96.2%	[4]
7	采用 Relief – F 进行基因初选,然后 采用 HykGene 进行精选	朴素 Bayes 分类器	3	91.9%	[19]
8	模糊逻辑与遗传算法 (Fuzzy Logic and GA)	SVM	10	99.41%	[20]
9	秩和检验方法 启发式宽度优先搜索(HBSA)	SVM	4	100%	[6]

2.4　相关工作的比较

由于基因表达谱的肿瘤分类检测对于医学临床诊断的重要性,肿瘤基因分类问题受到广泛关注。表 5 给出了针对结肠癌数据集采用不用基因选择方法或特征抽取方法的分类结果比较,这些都是目前肿瘤分类问题研究中获得的比较好的实验结果。比较发现,本节的结果在信息基因数量和识别精度方面具有明显的优越性,表明本文提出的 SVMMIV 算法与其它基因选择方法相比可以获得更好的效果。

3　讨论

癌症是由于正常组织在物理或化学致癌物诱导下,某些基因表达调控异常而导致的,因此了解基因的特异性表达与癌症之间的关系对于掌握癌症的发生和发展机制具有重要意义。随着大规模基因表达谱技术的发展,人们已经获得了人类各种组织正常基因和某些癌症基因的表达数据,因此对基因表达数据的分析与建模已经成为生物信息学研究的一个重要课题。每一种肿瘤都有其各自的基因特征表达谱,如果能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,对于诊断和治疗肿瘤将具有重要的意义。能否正确识别肿瘤类型、给出可靠诊断和简化实验分析,关键在于是否能从 DNA 芯片所测量的成千上万个基因中找出决定样本类别的基因"标签",即"信息基因"。

与其他公开发布的肿瘤数据集相比,结肠癌数据集比较难分类,绝大多数分类方法都很难获得 100% 的交叉验证准确率。本文以结肠癌特征基因选取为例,针对二类别肿瘤样本的分类问题,提出了基于支持向量机的基因平均影响值分析方法。该方法首先以 Wilcoxon 秩和统计量作为基因对肿瘤的辨识性度量进行无关基因滤除。然后以平均影响值分析为核心进行肿瘤特征基因筛选。从特征选取的角度看,这是一种基于 SVM 的 Filter – Wrapper 混合方法。采用 Filter 方法进行特征基因初选可以大幅降低特征变量的搜索空间、减少计算所需时间,而用 Wrapper 方法进行特征基因精选可以得到具有较好分类性能的特征变量。实验结果表明,该方法能很好地完成肿瘤分类特征选取的工作,只需要 3 个基因就能获得 100% 留一交叉验证分类正确率,为肿瘤基因表达数据的分析提供一种可以借鉴的方法,可望在实际肿瘤临床诊断和药物研制中得以应用。此外,通过查询相关生物医学文献发现本文获得的一些特征基因与肿瘤的发生发展有密切联系。例如,基因 R87126 (myosin heavy chain, nonmusle (Gallus gallus))是细胞组织中不可缺少的骨架蛋白,在肿瘤转移和细胞运动中起着重要作用;基因 K03460(Human alpha – tubulin isotype H2 – alpha gene, last exon.)在细胞有丝分裂和染色体分离中起关键作用,α 微管蛋白(α – tubulin)是细胞骨架的组成成分之一,有研究表明通过药物作用微管蛋白能够阻止肿瘤细胞的有丝分裂导致细胞进入凋亡期,从而抑制肿瘤生长,因此微管蛋白在肿瘤的发生和发展中起重要作用[21];基因 H23544(GTP – binding nuclear protein RAN (Homo sapiens))是一种分布于真核细胞核内

含量十分丰富的小分子 GTP 酶,是 Ras 基因大家族中的一员,由 Drivers 等[22]首次在人畸胎瘤 cDNA 中发现的第一个被鉴定为与人类肿瘤发生相关的蛋白,并且是一个基因表达调控子;基因 H06524 (gelsolin precursor, plasma (human))在几种癌细胞中都有表达的[23],凝溶胶蛋白(Gelsolin)是一种控制细胞凋亡的多功能肌动蛋白。这些实验结果从生物学上支持了本文算法的有效性。

参考文献(References)

[1] Vapnik V N. Statistical learning theory[M]. New York:Wiley Interscience,1998.

[2] Brown M, Grundy W N, Lin D, Cristianini N, Sugnet C W., Furey T S, Ares M Jr., Haussler D. Knowledge - based analysis of microarray gene expression data by using support vector machines[J]. Proc Natl Acad Sci, 2000, 97(1): 262 - 267.

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

[4] 邓林,马尽文,裴健. 秩和基因选取方法及其在肿瘤诊断中的应用[J]. 科学通报, 2004, 49(13): 1311 - 1316.

[5] 王星. 非参数统计[M]. 北京:清华大学出版社, 2009, 105 - 108.

[6] 王树林, 王戟, 陈火旺, 李树涛, 张波云. 肿瘤信息基因启发式宽度优先搜索算法研究[J]. 计算机学报, 2008, 31(4):636 - 649.

[7] Wang S L, Wang J, Chen H W, Tang W S. The Classification of Tumor Using Gene Expression Profile Based on Support Vector Machines and Factor Analysis. Intelligent Systems Design and Applications, Jinan, China[J]. IEEE Computer Society Press, 2006, 2: 471 - 476.

[8] Wang S L, Chen H W, Wang J, Zhang D X, and Li S T.. Molecular Diagnosis of Tumor Based on Independent Component Analysis and Support Vector Machines[C]. Proceedings of the 2006 international conference on computational intelligence and security, 2006, 1: 362 - 367.

[9] 黄德双, 刘海燕, 施蕴渝, 陈国良. 生物信息学中的智能计算理论与方法研究[M]. 合肥: 中国科学技术大学出版社, 2006.

[10] John G, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Cohen W W, Hirsh H, Eds. The Eleventh International Conference on Machine Learning[M]. San Fransisco: Morgan Kaufmann, 1994.

[11] 史峰,王小川,郁磊, 李洋. MATLAB 神经网络 30 个案例分析[M]. 北京:北京航空航天大学出版社,2010.

[12] Cristianini N, Shawe - Taylor J. An Introduction to Support Vector Machines[M]. Cambridge University Press, Cambridge, UK, 2000.

[13] Chang C C, Lin C J. LIBSVM: A library for support vector machines[EB/OL], Software available at http://www. csie. ntu. edu. tw/ ~ cjlin/.

[14] Breiman L, Spector P. Submodel selection and evaluation in regression: the X - random case[J]. International Statistical Review, 1992, 60(3): 291 - 319.

[15] Furey T S, Cristianini N, Duffy N Bednarski D W, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10): 909 - 914.

[16] Guyou I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machine[J]. Machine Learning, 2002, 46(1 - 3): 389 - 422.

[17] Peng S, Xu Q, Ling X B, Peng X, Du W, Chen L. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines[J]. FEBS Letter, 2003, 555(2): 358 - 362.

[18] Deb K, Reddy A R. Reliable classification of two - class cancer data using evolutionary algorithms[J]. BioSystems, 2003, 72: 111 - 129.

[19] Wang Y, Makedon F, Ford J C, Pearlman J. Hykgene:a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data[J]. Bioinformatics, 2005, 21(8): 1530 - 1537.

[20] Huerta E B, Duval B, Hao J K. A hybrid GA/SVM approach for gene selection and classification of microarray data[J]. Evo Workshops, 2006, 34 - 44.

[21] 刘青松, 张科伟, 薛梦华, 李智, 刘伟. α 微管蛋白在非小细胞肺癌中的表达及临床意义[J]. 中国实验诊断学, 2012, 16(3):447 - 451.

[22] Drivas G T, Shih A, Coutavas E, Rush M G, D'Eustachio P. Characterization of four novel ras - like genes expressed in a human teratocarcinoma cell line[J]. Mol Cell Biol, 1990,10: 1793.

[23] Kwiatkowski D J. Functions if gelsolin: Motility, signaling, apoptosis, cancer[J]. Current Opinion in Cell Biology, 1999, 11(1): 103 - 108.