文章编号:1000－2367(2017)02－0087－06 DOI:10.16366/j.cnki.1000－2367.2017.02.017

基于信噪比与随机森林的肿瘤特征基因选择

徐久成,冯森,穆辉宇

(河南师范大学 计算机与信息工程学院;河南省高校计算智能与数据挖掘
工程技术研究中心,河南 新乡 453007)

摘 要:在肿瘤特征基因选择过程中,传统分类方法会选出大量冗余基因,而大量冗余基因会造成分类精度低和时间复杂度较高等问题,为了解决上述问题,提出一种结合信噪比过滤法与随机森林算法的肿瘤特征基因选择方法.该方法包含两个过程:首先使用信噪比过滤法剔除原始特征空间中的无关和冗余基因,从而获得与分类属性相关性较高的基因,选择出分类能力较强的预选特征子集;其次使用随机森林算法对特征基因子集进行分类,最终获得分类结果.实验结果显示,该算法可以快速有效地选择出肿瘤特征基因,并具有较高的分类精度.

关键词:基因表达谱;特征选择;信噪比;随机森林

中图分类号:TP181 **文献标志码**:A

近年来,伴随 DNA 芯片技术的快速成长与成熟[1]以及基因表达谱数据(GEP)的成倍增长,这些为人类身体状况和疾病分析与辨别提供了有效的帮助[2].GEP 通常包含几千个甚至上万个基因,通常仅有几十个样本,而且实际上只有少数基因与分类有关,这些基因被称为信息基因,又称为特征基因[2－4],由于肿瘤基因表达谱数据具有高维数、高噪声、高冗余等特点,因此如何从基因数据中选择出对疾病有辨别意义的特征基因或与疾病相关的特征基因成为生物信息学的研究热点之一.

特征基因选择的目标是减少噪音数据和冗余数据,从基因表达谱数据中选择特征基因子集,从而使得到的特征基因子集具有较强的疾病辨别能力[5－8].现在最常用的特征基因选择有两种方法:过滤法(Filter)和缠绕法(Wrapper)[9],本文所用的信噪比指标是属于过滤法,其优势是运算速度快,它根据基因对样本分类的贡献大小,对无关基因剔除和过滤,从而提高肿瘤基因的识别概率;然而它的不足是没有全面的考虑到基因相互之间的关联度,导致分类精度不高[10].

为了获得更好的分类效果,使用现阶段对高维数据分类与回归有着卓越性能的算法—随机森林(Random Forest,简称 RF).随机森林算法是集成的机器学习算法[11],同时也是 Bagging 的一个扩展变体.首先使用随机重抽样方法 Bootstrap 和节点随机分裂方法生成多棵决策树,并在以决策树为基学习器构建 Bagging 集成的基础上,进一步在决策树的训练过程中引入了随机属性选择,然后采用投票的方式获得分类结果.RF 具有分析复杂相互作用分类特征的能力,对存在缺失值的数据具有很好的鲁棒性,并且学习速度非常快,其特征重要性度量可以用来对高维数据进行特征选择,近年来该算法已广泛应用于各种数据分类、蛋白质预测、特征选择以及异常点检测问题中[12].文献[13]将随机森林与 K 近邻算法相结合,提出了一种新的投票机制,充分利用决策树上的 OOB(Out-of-Bag)信息,较好地提高分类精度;文献[14]提出一种应用随机森林算法进行特征选择的方法,以随机森林分类精度为准则函数对特征进行重要性度量的方式实现特征选择;以上这些算法都表现出了良好的分类性能,但仍然存在分类结果不稳定和冗余基因过多造成的子集规模过大等问题,如何在保证分类性能的前提下,提高算法稳定性和减小特征子集规模是肿瘤基因特征选择的重要研究

收稿日期:2016-11-14;**修回日期**:2017-02-21.

基金项目:国家自然科学基金(61370169;61402153);河南省科技攻关重点项目(142102210056;162102210261).

作者简介:徐久成(1964－),男,河南偃师人,河南师范大学教授,博士,博士生导师,研究方向为粒计算、数据挖掘、粗糙集、生物信息学等.

通信作者:冯森,E-mail: fengsen@htu.edu.cn.

问题.

鉴于肿瘤基因表达谱数据的高维数、小样本、噪声冗余基因多而有用信息基因少等特点[15],本文结合了信噪比特征提取方法的运算速度快和随机森林简单易实现、计算时间短、分类精度高等优点,提出了一种基于信噪比与随机森林的肿瘤特征基因选择方法.该方法主要包含两个步骤:首先使用信噪比过滤法对基因数据进行处理,接着选择出预选特征子集;然后采用随机森林算法对预选子集进行分类;通过实验分析,本文的算法明显优于已有的特征基因选择算法,并且在分类性能上也有较大提高.

1 基本理论

1.1 信噪比

在基因表达谱数据研究中,信噪比指标作为基因选择准则,是一种非常简单且高效的基因排序准则[10].在进行特征基因选择时,首先利用信噪比指标对原始基因数据集计算每个基因对分类属性的关联度,然后进行排序,从而过滤和剔除掉冗余基因,最后获得与分类属性关联度较高的基因子集,信噪比的计算公式为

$$S(g_i)=\frac{|u_+(g_i)-u_-(g_i)|}{\delta_+(g_i)+\delta_-(g_i)},\tag{1}$$

式中:$u_+(g_i)$ 和 $u_-(g_i)$ 代表第 i 个基因 g_i 在正类和负类对应的均值;而 $\delta_+(g_i)$ 和 $\delta_-(g_i)$ 代表第 i 个基因 g_i 在正类和负类对应的方差.(1) 式作为每个基因与分类属性关联度的判别标准,(1) 式的值越大,说明该基因与分类属性的关联度越大.

1.2 随机森林

随机森林算法是 2001 年由 Leo Breiman 和 Adele Cutler 两人共同提出的一种集成的机器学习算法,它集成了由 Breiman 提出的 Bagging 算法,CART(Classification And Regression Tree)决策树思想和贝尔实验室的 Tin Kam Ho 与 Geman 提出的随机选择思想[16].下面给出随机森林相关定义与定理.

定义 1 随机森林[11]是一个由多个决策树分类器组成的组合分类器 $\{h(X,\theta_k),k=1,2,\cdots,K\}$,其中 X 表示自变量,参数 $\{\theta_k\}$ 表示服从独立同分布的随机向量,K 表示决策树分类器的个数.在给出 X 的情况下,使用 bootstrap 重抽样的方法,构建决策树模型并对其进行训练,可以把决策树中的每一棵树都看成是一个弱分类器,通过对每个弱分类器的分类结果,以投票的方式获得最优的分类结果.

定义 2 给定一组决策树分类器 $h_1(X),h_2(X),\cdots h_k(X)$,分类器所使用的训练集都是从原始随机向量 (Y,X) 中随机获得,间隔函数(Margin function)定义为 $mg(X,Y)=av_kI(h_k(X)=Y)-\max\limits_{j\neq Y}av_kI(h_k(X)=j)$,式中:$I(\cdot)$ 是指示函数,间隔函数用于度量平均分类数超过平均错误分类数的程度,间隔函数的值越大,分类结果越可信.泛化误差表示为 $PE^*=P_{X,Y}(mg(X,Y)<0)$ 其中,下标 X,Y 表示概率 P 在空间 X,Y 上.

在随机森林中,在决策树的个数足够多的情况下,$h_k(X)=h(X,\theta_k)$ 服从强大数定律.

定理 1 随着随机森林中决策树数的增加,所有序列 $\theta_1,\theta_2,\cdots,\theta_k,PE^*$ 几乎处处收敛于 $P_{X,Y}\{P_\theta(h(X,\theta)=Y)-\max\limits_{j\neq Y}P_\theta(h(X,\theta)=j)<0\}$,由定理 1 可知,随着决策树数目的增加,随机森林没有产生过拟合的问题,但在一定程度上产生了泛化误差.

随机森林算法的一个最重要的特点就是采用变量重要性评估.一般情况下,随机森林程序会给出 4 种变量重要性度量方法.本文采用的是基于袋外数据(Out-of-Bag)分类准确率的变量重要性度量.

定义 3 基于袋外数据(Out-of-Bag)分类准确率的变量重要性度量定义为袋外数据自变量值发生轻微扰动后分类正确率与扰动前分类正确率的平均减少量[17],重要性度量公式为:$D=\frac{1}{B}\sum\limits_{i-1}^{B}(R_b^o-R_{bj}^o)$.

定义 4 随机森林算法的分类准确率定义为[18]:$A=\frac{T_P+T_N}{T_P+T_N+F_P+F_N}$,式中:$T_P$(true positive)代表正确的肯定;$T_N$(true negative)代表正确的否定;$F_P$(false positive)代表错误的肯定;$F_N$(false negative)代表错误的否定.

2　特征基因选择算法

2.1　数据预处理

本文对原始基因数据集使用信噪比过滤法进行预处理，根据信噪比值的大小以升序的方式对全部基因数据进行排序，由于信噪比的值域是(0,1]，所以以 0.2 为单位把已排序的基因序列划分成 5 个区间，划分的区间依次为(0,0.2]，(0.2,0.4]，(0.4,0.6]，(0.6,0.8]，(0.8,1]. 以上 5 个区间就是特征基因子集，每个区间中的数据都可以用作预选特征子集. 通过(1)式计算得知，基因的信噪比值越大，说明该基因与分类属性的关联度就越大.

本文为了选择出噪声数据少且与分类属性相关性高的预选特征基因子集，只选择信噪比值较大的基因作为预选特征基因. 由于 4 个数据集的信噪比值在区间(0.8,1]的特征基因数目几乎都为零，所以为了提高算法的整体性能，将信噪比值在区间(0.8,1]内的基因作无关基因处理. 最终选取各个数据集在区间(0.6,0.8]内的基因作为预选特征基因子集.

2.2　基于信噪比与随机森林的肿瘤特征基因选择算法

本文提出了基于信噪比与随机森林的肿瘤特征基因选择算法，首先通过计算每个基因的信噪比值的大小，对原始基因数据进行升序排序，根据排序结果过滤掉无关基因，选择出与分类属性相关度较高的预选特征子集，其次采用随机森林算法进行分类，利用 bootstrap 重抽样方法，构建一组由多棵决策树组成的分量分类器，决策树中的每个节点都可以看作是一个弱分类器，然后利用投票机制(voting)对分类结果进行投票，最终获得基因个数最少、分类准确率最高的特征基因集合作为特征选择的结果. 为了确保实验结果的稳定性及可靠性，故采用多次实验求平均值的方法在数据集上进行验证. 将 10 次实验结果的分类正确率求平均值，作为本文算法的最终分类正确率. 该算法描述如算法 1 所示.

算法 1 基于信噪比与随机森林的肿瘤特征选择

输入：原始基因数据集 $S_0=(x_1,x_2,\cdots,y)$. 输出：特征基因集合 S.

步骤 1　对原始基因数据集 S_0 进行预处理，即将缺值设置为 0；

步骤 2　根据(1) 式对每个基因的信噪比值进行计算；然后根据信噪比值的大小对 G_i 进行升序排序；//G_i 代表经过信噪比排序后的基因序列；

步骤 3　将排好顺序的基因序列 G_i 通过信噪比去除冗余基因，获得与分类属性关联度较高的预选特征基因子集；

步骤 4　采用 Bootstrap 方法进行有放回地随机抽取 k 个样本集，k 取值为原始样本数量的 2/3；

步骤 5　抽到的每个样本集为决策树生长的训练集，每次未被抽到的样本组成 h 个袋外数据；

步骤 6　随机森林利用 Bagging 方法与 Bootstrap 方法两种重采样技术生成多个决策树基分类器；

步骤 7　使用步骤 6 所构建的基分类器在测试集上对测试数据集进行预测分类；

步骤 8　标记各个决策树的最终分类结果，计算全部决策树对不同类标签的投票结果；

步骤 9　分析投票结果，采取少数服从多数的原理，投票数多的所对应的类标签为最终的样本类别；

步骤 10　分别计算算法在测试样本集上的误分率和运行时间，最终获得特征基因子集 S；

步骤 11　结束.

2.3　时间复杂度分析

本文先采用信噪比对数据集过滤掉无关基因，然后通过随机森林算法进行特征选择. 根据(1)式可以得到信噪比的时间复杂度为基因的样本数 n. 随机森林的基分类器是分类回归树(Classification And Regression Tree，CART)，分类回归树是一棵二叉树. 若训练集的特征数为 m，样本为 n，CART 算法的时间复杂度为 $O(mn(\lg n)^2)$. 随机森林在构建 CART 树的过程中，从 m 个特征中随机选择 m_i 个特征计算信息增益，并且对树的生长不进行剪枝，故训练每一个基分类器的计算时间小于 $O(mn(\lg n)^2)$，设随机森林中基分类器的个数为 k 个，则随机森林算法的时间复杂度可以近似为 $O(kmn(\lg n)^2)$[18]，故本文的时间复杂度为 $O(n)+O(kmn(\lg n)^2)$. $O(n)$ 相对于 $O(kmn(\lg n)^2)$ 的复杂度可以忽略，最终的时间复杂度约为 $O(kmn(\lg n)^2)$.

3 实验结果与分析

3.1 实验数据与方法

为了验证算法的有效性,本文在 Lung,Colon,Leukemia 和 Prostate 4 个公开的数据集上进行实验仿真,本文实验所用到的数据集是从 http://datam.i2r.a-star.e-du.sg/datasets/krbd/下载获得,具体的数据集描述见表 1[10].仿真实验用到的计算机配置为酷睿 i3-2350M,2.30 GHz,8 GB 内存,所有仿真都在 MATLAB R2013a 中实现,并与 ODP,SNRS,RF,SVM 4 种分类方法进行对比.

表 1 实验数据集描述

序 号	数据集名称	基因数量	样本数量(正类/负类)	类别数
1	Lung	2880	39(15/24)	2
2	Colon	2000	62(40/22)	2
3	Leukemia	7129	72(25/47)	2
4	Prostate	12 600	102(52/50)	2

3.2 实验结果分析

通过计算信噪比值的大小,发现 4 个数据集的信噪比值在区间(0.8,1]的特征基因数量几乎都为零,故将基因特征分布在 4 个区间 (0,0.2],(0.2,0.4],(0.4,0.6],(0.6,0.8],数据集在对应区间的基因数目依次是{1727,969,174,10},{1364,564,72,0},{4973,1796,334,26},{6976,5156,411,49}.通过 MATLAB 实验仿真得知大部分基因的信噪比值都比较小,如 Prostate 数据集的基因个数为 12 600,并且有 6976 个基因的信噪比值小于或者等于 0.2;Leukemia 数据集的基因个数为 7129,但是有 4973 个基因的信噪比值小于或者等于 0.2.由于低信噪比的特征基因很难达到区分类别的作用,可视为无关基因,仅有少数的基因与样本的分类属性高度相关.为了更好地获取最优特征基因子集和分类性能,本文只把比值在区间(0.6,0.8]内的基因作为候选的特征基因子集.又由于 Colon 数据集在区间(0.6,0.8]的基因个数为零,所以选择区间(0.4,0.6]中的基因作为特征基因子集.因此通过对以上 4 个数据集采用信噪比过滤掉无关基因之后,最终得到预选的特征基因子集数量分别是 10、72、26 和 49.

在随机森林算法中,计算每个样本作为 OOB 样本的树对它的分类情况,然后通过简单多数投票的方式得到该样本的分类结果,最后将误分个数占样本总数的比值作为的 OOB 误分率,图 1 至图 4 为 4 个数据集的在不同个数的决策树下的 OOB 误分率的变化趋势.从图 1 至图 4 可以看出随着随机森林中决策树个数的递增,该算法既不会产生过拟合,收敛速度也明显较快,且结果的误分率较低.

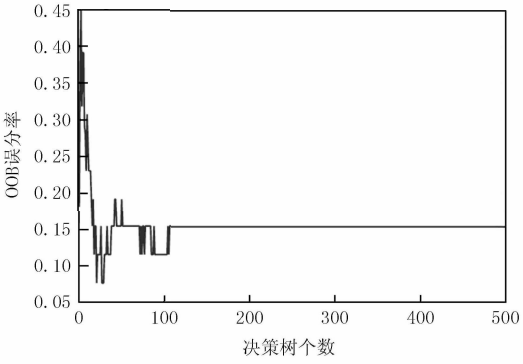

图1 Colon数据集上的OOB误分率

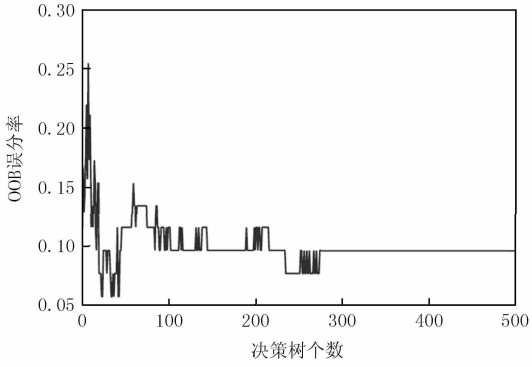

图2 Lung数据集上的OOB误分率

图 5 给出了 5 种分类算法在 4 个数据集上的分类性能的对比,ODP(Original data processing)表示对原始数据集直接进行分类的方法; SNRS 表示为采用基于信噪比与邻域粗糙集对原始数据集进行分类的方法;RF 表示为只采用随机森林方法进行分类的方法;SVM 表示为采用支持向量机进行分类的方法;SNRRF 表示为本文算法采用基于信噪比与随机森林进行分类的方法.由图 5 可知,对 4 个数据集分别使用不同的分

类算法，得到了不同的分类效果. 通过对比不同分类算法的分类性能，本文提出的基于信噪比与随机森林算法的分类正确率相对较高.

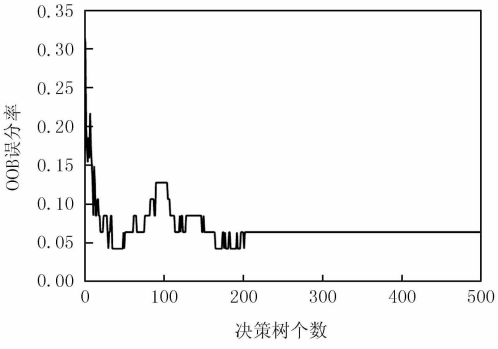

图3　Leukemia数据集上的OOB误分率

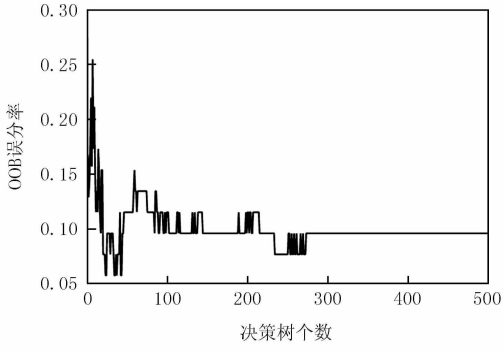

图4　Prostate数据集上的OOB误分率

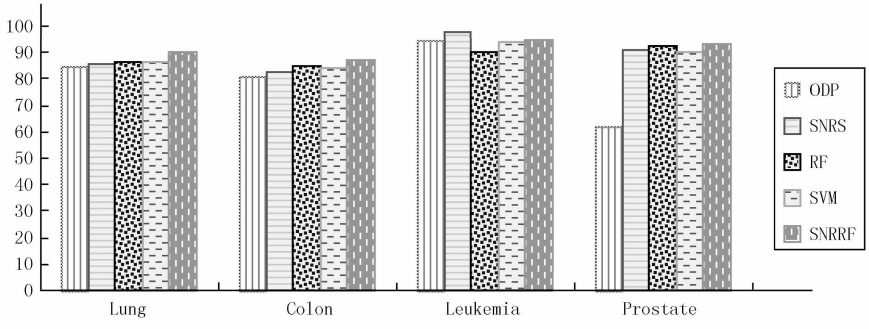

图5　4个数据集用不同算法的分类清度

根据表 2 可以看出，在 Lung 数据集中，本文方法的分类正确率为 89.89%，在 Colon 数据集中，本文算法的分类正确率为 87.48%，在 Prostate 数据集中本文算法的分类正确率为 93.14%，均高于其他 4 种算法的分类正确率；但在 Leukemia 数据集中，本文算法的分类正确率为 94.77%，略低于 SNRS 算法在数据集上 97.36%的分类正确率，这说明本文采用信噪比进行无关基因过滤时，错误的过滤掉了对分类影响较大的特征基因，从而影响了分类正确率. 以上算法的时间复杂度分别为 $O(mn)$，$O(mn\lg n)$，$O(kmn(\lg n)^2)$，$O(n^2)$，$O(kmn(\lg n)^2)$. 实验结果表明，本文算法在其他 3 个数据集上都表现出了非常好的分类效果，不仅能够选择出关联度高、低冗余度的特征基因子集，而且有效地提高了特征分类算法的正确率.

表 2　各种算法在不同数据集上的特征基因个数和分类性能

方　法	Lung		Colon		Leukemia		Prostate		时间复杂度
	基因数	分类性能/%	基因数	分类性能/%	基因数	分类性能/%	基因数	分类性能/%	
ODP	2880	84.62	2000	81.10	7129	94.44	12 600	61.90	$O(mn)$
SNRS	6	85.44	15	82.26	4	97.36	5	91.18	$O(mn\lg n)$
RF	2880	86.37	2000	84.75	7129	90.18	12 600	92.54	$O(kmn(\lg n)^2)$
SVM	16	86.36	15	84.40	10	94.10	10	90.34	$O(n^2)$
SNRRF	10	89.89	72	87.48	26	94.77	49	93.14	$O(kmn(\lg n)^2)$

通过表 2 可知，虽然 ODP 与 RF 两种算法在数据集上都可以获得较高的分类正确率，但是得到的基因子集规模过于庞大，造成算法运行速度慢，时间复杂度较高；SNRS 算法既获得了基因数目相对较少的特征子集，又提高了分类精度.

然而评价一个特征选择算法的优劣不仅要获得尽可能少的特征基因子集，还应该具有较高的分类正确率，SVM 算法的分类精度较高，时间复杂度较小，但是确定最优参数值耗时多. 本文算法的分类精度相对于其他 4 种分类算法最高，基因数目也相对最少，时间复杂度相对较低.

4 结 论

为了提高肿瘤特征基因选择的准确性,本文结合了过滤法和集成机器学习方法,提出了一种基于信噪比与随机森林的肿瘤特征基因选择方法.该方法能够快速高效地从原始特征空间中得到特征基因子集,并且能获得更高的分类精度.通过实验仿真与其他算法比较,本文算法能够选择出与肿瘤相关度高并且规模较小的特征基因子集,并提高了分类正确率.如何进一步减少特征基因个数以及降低随机森林算法的泛化误差,将是下一阶段的研究内容.

参 考 文 献

[1] Golub T R, Slonim D K, Tamayo P, et al. Class Discovery and Class Prediction by Gene Expression Monitoring[J]. Brain Research, 1999, 501(2): 205-14.

[2] 明利特,蒋芸,王勇,等.基于邻域粗糙集和概率神经网络集成的基因表达谱分类方法[J].计算机应用研究,2011,28(12):4440-4444.

[3] 李颖新,李建更,阮晓钢.癌症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(12):324-330.

[4] SUN L, XU J C, REN J Y, et al. Granularity Partition-based Feature Selection and its Application in Decision Systems[J]. Journal of Information and Computational Science, 2012, (12): 3487-3500.

[5] 段艳华.基于基因表达谱的肿瘤分类特征基因选择研究[D].北京:北京工业大学,2008.

[6] 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.

[7] 谢娟英,胡秋锋,董亚非.K-S 检验与 mRMR 相结合的基因选择算法[J].计算机应用研究,2016,04:1013-1018.

[8] 孙伟,韩飞.基于基因灵敏度信息和二进制微粒群优化的基因选择方法[J].计算机应用研究,2014,09:2648-2651.

[9] 关健,韩飞,杨善秀.基于粒子群优化和判别熵信息的基因选择算法[J].计算机工程,2013,39(11):187-190.

[10] 徐久成,李涛,孙林,等.基于信噪比与邻域粗糙集的特征基因选择方法[J].数据采集与处理,2015,30(5):973-981.

[11] Breiman L. Random Rorests[J]. Machine Learning, 2001, 45(1): 5-32.

[12] Strobl C, Boulesteix A L, Kneib T, et al. Conditional Variable Importance for Random Forests[J]. BMC Bioinformatics, 2008, 9(1): 1-11.

[13] 杨帆,林琛,周绮凤,等.基于随机森林的潜在 k 近邻算法及其在基因表达数据分类中的应用[J].系统工程理论与实践,2012,32(4):815-825.

[14] 袁晓龙,梅雪,黄嘉爽,等.基于随机森林算法的特征选择及在 fMRI 数据中的应用[J].微电子学与计算机,2014(8):132-135.

[15] 徐久成,徐天贺,孙林,等.基于邻域粗糙集和粒子群优化的肿瘤分类特征基因选取[J].小型微型计算机系统,2014(11):2528-2532.

[16] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[17] Verikas A, Gelzinis A, Bacauskiene M. Mining Data with Random Forests: A Survey and Results of New Tests[J]. Pattern Recognition, 2011, 44(2): 330-349.

[18] 姚登举,杨静,詹晓娟.基于随机森林的特征选择算法[J].吉林大学学报(工学版),2014,44(1):137-141.

Tumor Feature Gene Selection Based on SNR and Random Forest

Xu Jiucheng, Feng Sen, Mu Huiyu

(College of Computer & Information Engineering; Henan Engineering Technology Research Center for Computing Intelligence & Data Mining; Henan Normal University, Xinxiang 453007, China)

Abstract: Given in the process of tumor feature gene selection, the traditional classification methods selected a large number of redundant genes, which led to a lower classification precision and higher time complexity. In order to solve the above-mentioned problems, this paper proposed a tumor gene feature selection method based on Signal Noise Ration and Random Forest. The method includes two processes: firstly, it filtered the irrelevant genes in the original feature space using the index of signal noise ratio, and obtained the genes which were closely related to the categorical attributes, then chosen the primary character subsets with higher capability of classification; secondly, classify the obtained character subsets with the random forest algorithm, finally the classification results were obtained. The experimental results show that the proposed method not only quickly and efficiently selected feature gene but also has a higher classification precision.

Keywords: gene expression profiles; feature selection; signal-to-noise ratio; random forest

[责任编辑 陈留院]