文章编号:1000－2367(2018)02－0104－07 **DOI**:10.16366/j.cnki.1000－2367.2018.02.017

基于 PCA 和信息增益的肿瘤特征基因选择方法

徐久成, 黄方舟, 穆辉宇, 王云, 徐战威

(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:针对肿瘤基因数据因维度高和冗余基因较多而导致分类精度低的问题,提出一种基于 PCA 和信息增益的肿瘤特征基因选择方法.该方法首先使用 PCA 算法剔除冗余基因,获得预选特征基因子集;然后利用信息增益算法对预选特征基因子集进行优化选取,得到特征基因子集;最后采用不同分类模型对特征基因子集进行仿真实验.实验结果表明,所提方法提高了基因表达谱的分类精度,从而表明致病基因被有效地选取出来.

关键词:基因分类;主成分分析;信息增益;特征选择

中图分类号:TP181 **文献标志码**:A

近年来,国内外癌症的高发率和死亡率日渐显著,成为人类亟待解决的问题,而与癌症相关的肿瘤基因只存在于极少的表达基因序列中[1].为及时发现并有效预防癌症,越来越多的研究者致力于研究肿瘤基因表达谱数据集.特征基因选择即从基因表达谱数据集中获取具有较强癌症识别能力的特征基因子集[2].经过大量实验验证:影响肿瘤基因分类精度的关键问题在于特征基因的选择方法[3].因此肿瘤特征基因选择方法逐渐成为研究肿瘤发病原理及临床疾病诊断的热门技术之一[4－5].

传统的分类算法,往往存在着维数灾难的问题,其原因是传统的分类算法未对数据集进行预选处理,而是直接对其构建的矩阵进行分析,因此传统的分类算法存在着较高的时间复杂度和较低的分类精准度等问题[6].伴随着 DNA 表达微阵列技术越来越成熟,研究者针对肿瘤基因表达数据的样本小、维数高和噪声大等特点,做了大量有关特征基因选择方法的研究.文献[7]提出 T-test 测试方法,有效地提取出与白鼠缺乏高脂蛋白相关的基因.文献[8]将 Fisher 准则和多类相关矩阵结合,采用评价函数计算并获得最优特征基因子集,有效地剔除冗余基因.文献[9]采用 ReliefF 算法及改进的邻域粗糙集模型提高分类精度且减少时间复杂度.文献[10]将邻域互信息最大化和粒子群优化算法巧妙结合,快速有效地进行特征基因选择.文献[11]采用 Pearson 相关系数和 Wilcoxon 秩相结合的方法,选取预选特征基因子集,然后通过 SVM 分类器对预选特征基因子集进行分类.文献[12]将适应度函数引入特征基因选择方法,通过计算定义类间和类内间距的比值提高肿瘤基因的分类精度 .文献[13]提出一种结合随机森林和邻域粗糙集的特征选择方法,实验证明该特征基因选择方法不仅能有效选取特征基因子集且实验分类性能良好.文献[14]将信噪比与随机森林结合,有效地提高肿瘤基因分类精度并降低实验的时间复杂度.文献[15]将稀疏主成分与 K-means 方法结合提高特征基因聚类的精准性和高效性.文献[16]将 Fisher 权函数、离散傅里叶和主成分分析结合,有效地提高结肠癌数据集的分类精度.但是以上这些方法普遍存在模型泛化能力差和分类精度偏低等问题.目前常用的特征基因选择方法有过滤法(Filter)、嵌入法(Embedded)和缠绕法(Wrapper)[17].过滤法具有快速简便的特点,其方法求得的值即基因对应的得分,然后对其得分进行排序,最后选取分值较高的(即重要度较高)特征基因集.信息增益方法有着非常广泛的应用领域.文献[18]提出一种改进的信息增益特征优化方法处理文本分类问题,其主要原理是将频度、分散度和集中度引入信息增益中.文献[19]对信息增益加权处理,并与 N-gram 结

收稿日期:2017-08-19;**修回日期**:2017-10-27.

基金项目:国家自然科学基金(61370169;60873104);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:黄方舟,E-mail:hfz.htu@foxmail.com.

合，有效地提高恶意代码的检测率和准确率.文献[20]通过对样本加权处理，从而改进信息增益方法，构建特征基因选取模型并验证模型的稳定性.

本文结合主成分分析和信息增益算法进行研究，构建肿瘤特征基因选择模型.考虑保留有效特征基因、时间复杂度及算法稳定性等因素，首先采用主成分分析(Principal Component Analysis，PCA)算法对肿瘤基因数据集进行降维处理从而获得预选特征基因数据集；然后采用信息增益算法对预选特征基因数据集进行数据打分，选取出分值较高的基因从而达到优化预选特征基因数据的目的；最后使用 Weka 工具中的几种不同分类算法进行分类验证.仿真实验表明本文所用方法可以提高分类精度，从而验证了该方法的有效性.

1　基本概念

1.1　主成分分析

主成分分析算法(简称 PCA 算法)是一种通过降低数据集维度，从而达到有效提取特征数据子集目的的算法.1901 年 Karl Parson 首次提出主成分概念，1933 年此概念被 Hotelling 应用到随机变量研究方面[21].该算法思想是通过 PCA 算法处理数据集，分析其主成分从而获得特征的贡献率.将特征根据贡献率作降序处理，选取贡献率较大的特征，构建特征子集.

定义 1[22]　设一数据集有 K 个样本，每个样本包含 M 个特征.把 K 个样本和 M 个特征分别从左向右和从上向下排列，得到一个 M 行、K 列的矩阵 X，矩阵 X 中第 i 列表示第 i 个样本 $i=(1,2,\cdots,K)$，则有 $X=(x_1,x_2,\cdots,x_K)$.

计算平均样本 $\overline{X}$，

$$\overline{X}=\frac{1}{K}\sum_{i=1}^{K}x_i. \tag{1}$$

计算差值样本 d_i，

$$d_i=x_i-\overline{X},i=1,2,\cdots,K, \tag{2}$$

构建协方差矩阵 C，

$$C=\frac{1}{K}\sum_{i=1}^{K}d_id_i^{\mathrm{T}}=\frac{1}{K}AA^{\mathrm{T}},A=(d_1,d_2,\cdots,d_K), \tag{3}$$

AA^{T} 为 M 阶方阵，L 为 M 维非零向量，若

$$AA^{\mathrm{T}}L=\lambda L, \tag{4}$$

则 λ 和 L 分别为 AA^{T} 的特征值和特征向量.

贡献率是衡量每个特征携带有效信息的数值，λ_j 表示第 j 个特征的主成分方差，即相应的特征值，则第 j 个特征贡献率 μ_j，

$$\mu_j=\lambda_j/\sum_{j=1}^{M}\lambda_j,j=1,2,\cdots,M. \tag{5}$$

1.2　信息增益

信息增益(Information Gain，IG)是一种有效的特征选择方法，被广泛应用于机器学习领域，是用来解决特征选择问题较有效的算法.信息增益算法主要用于刻画每个特征的重要程度，从所得结果中选取部分信息增益较大的数据(即携带有效信息较多的特征)，从而达到消除冗余特征的目的[9].其中信息增益值越大，表示该特征越重要.

定义 2[23]　设数据集 D 中 K 个类 $C_k,k=1,2,\cdots,K$，$|C_k|$ 为属于类 C_k 的样本个数，$|D|$ 代表数据集中所有样本数量.设特征 A 有 n 个不同的取值 $\{a_1,a_2,\cdots,a_n\}$，根据特征 A 的取值将数据集 D 划分为 n 个子集，即 $D=(D_1,D_2,\cdots D_i,\cdots,D_n)$，其中 $|D_i|$ 代表第 i 类样本的个数，$|D_{ik}|$ 代表 D_i 中属于类 C_k 的样本的个数.数据集 D 的经验熵 $H(D)$，

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}. \tag{6}$$

根据数据集 D 的经验熵 $H(D)$，可得特征 A 关于数据集 D 的经验条件熵 $H(D|A)$，

$$H(D\mid A)=\sum_{i=1}^{n}\frac{\mid D_i\mid}{\mid D\mid}H(D_i)=-\sum_{i=1}^{n}\frac{\mid D_i\mid}{\mid D\mid}\sum_{k=1}^{K}\frac{\mid D_{ik}\mid}{\mid D_i\mid}\log_2\frac{\mid D_{ik}\mid}{\mid D_i\mid}. \tag{7}$$

由(6)式和(7)式,可得信息增益 $I_G(D,A)$,

$$I_G(D,A)=H(D)-H(D\mid A). \tag{8}$$

2 基于 PCA 和信息增益的特征基因选择方法

2.1 基于 PCA 和信息增益的特征基因选择方法流程图

基于 PCA 和信息增益算法的工作流程为:首先将原始基因数据预处理,使得每一维均值都为 0,并构建其协方差矩阵,计算每个协方差矩阵的特征值和特征向量,可得每个基因相应的贡献率,选取贡献率大于 0.01的基因,从而得到降维后的预选特征基因矩阵,并对其进行归一化处理.计算低维特征基因矩阵的信息增益,并将得到的结果进行降序处理,选取信息增益值较大的部分基因,最后选择出较优特征基因子集 Y.此过程旨在去除冗余基因,尽可能多地保留携带有效信息的基因.基于 PCA 和信息增益的特征基因选择方法流程图如图 1 所示.

2.2 基于 PCA 和信息增益的肿瘤特征基因选择算法

针对基因表达谱数据高维度和低样本等问题,提出基于 PCA 和信息增益的肿瘤特征基因选择方法.首先采用 PCA 算法在 3 个标准的基因表达谱数据集(Lung、Colon 和 Leukemia)上计算每个数据集中各个基因的贡献率,由(5)式可知,基因的贡献率越大,则其携带的有效信息越多,并将贡献率由大到小排序,经过多次实验验证,选取贡献率大于 0.01 的基因构建出预选特征基因子集,从而达到剔除冗余基因的效果;然后采用信息增益算法计算预选特征集每个特征的信息增益,并比较信息增益的数值,选取部分信息增益较大的基因数据作为特征选择的结果.

算法 1 基于 PCA 和信息增益的肿瘤特征基因选择算法(PCA-IG)

输入 基因数据集 $X=(x_1,x_2,\cdots,x_K)$;

输出 特征基因集合 Y.

步骤 1 对原始基因数据集 X 进行预处理,构建其协方差矩阵 C;

步骤 2 计算矩阵 C 中特征向量 L 和特征值 λ;

步骤 3 根据贡献率(5)式,计算每个基因的贡献率 μ;

步骤 4 对得到的贡献率降序处理,并筛选出 $\mu>0.01$ 的基因;

步骤 5 筛选出的基因构建预选特征基因子集 D;

步骤 6 计算预选特征基因子集 D 的信息熵 $H(D)$;

步骤 7 根据(7)式可得特征 A 对于预选特征基因子集 D 的经验条件熵 $H(D\mid A)$;

步骤 8 由(8)式可得信息增益 I_G;

步骤 9 对信息增益 I_G 降序处理并筛选出特征基因集合 Y;

步骤 10 结束.

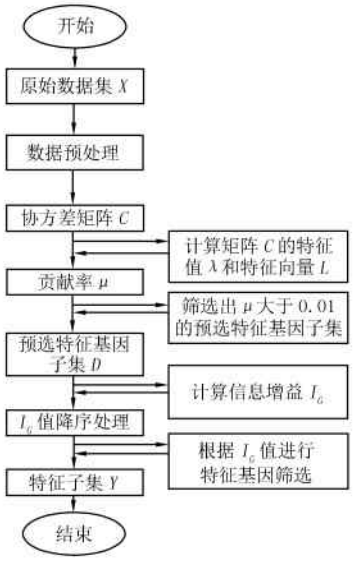

图 1 基于 PCA 和信息增益的特征基因选择方法流程图

3 实验分析

3.1 实验数据

文中仿真实验的对象是 3 种公开的二分类基因表达谱数据集(即 Lung、Colon 和 Leukemia),基因表达

谱数据集分别为肺癌基因数据集、结肠癌基因数据集和白血病基因数据集.其下载来源为：http://featureselection.asu.edu/datasets.php.该数据集详细介绍如表 1 所示.

表 1　数据集信息

序号	数据集名称	特征基因数量	样本类别	样本数量
1	Lung	12 600	Normal/Abnormal	202
2	Colon	2 000	Positive/Negative	61
3	Leukemia	7 129	ALL/AML	71

3.2　实验环境及平台

仿真实验环境及平台详细情况如表 2 所示.

表 2　实验环境及平台信息

实验环境	CPU　AMD Athlon II X4 645 Processor 内存　4 G 系统　Windows 7
实验平台	Matlab-R2010a Weka-3.9.0 SPSS

3.3　实验结果

PCA 算法首先通过分析 3 种基因表达谱数据集的主成分方差、主成分协方差及主成分得分等，根据普遍做法选取主成分得分(即各主成分贡献率)大于 0.01 的基因作为预选特征基因子集，从而达到对数据降维处理的目的，并得到相应的 3 个预选特征基因子集.然后采用信息增益计算预选特征基因子集的信息增益，并按降序的方法排列信息增益，选取信息增益值较大的基因，从而优化预选特征基因子集.本文的实验，在其他分类算法和本文所提出的 PCA-IG 算法中统一采用十折交叉验证，下面分别从所采用的方法在相同数据集和相同分类算法上与其他特征基因选择方法进行分类精度的对比、数据集经 PCA-IG 算法处理前后在不同分类算法上的分类精度对比和 PCA-IG 算法与其他基因选择方法最优分类精度对比 3 个方面，进行算法验证.

本实验首先在 C4.5、Naive Bayesian 和 LibSVM 3 种分类算法上与其他特征选择方法作分类性能的比较，采用的 PCA-IG 算法在所用参数较优情况下进行实验，如图 2～4 所示.

在图 2～4 中，ODP 表示直接对原始数据分类的方法；NRS 表示仅仅采用邻域粗糙集的方法；SNRS[24] 表示采用基于信噪比和邻域粗糙集的方法.由图 2～4 可知，针对相同数据集采用相同分类，本文的 PCA-IG 特征选择方法的分类性能相对较高.例如，在 Lung 数据集中，本文方法在 C4.5、Naive Bayesian 和 LibSVM 实验的分类精度分别为 97.54%、97.04%和 91.63%，分类精度明显高于其他分类方法.但是对于 Leukemia 数据集，本文方法在 LibSVM 上的分类精度低于 ODP 方法，表明本文方法在选择特征基因时，错误的删除对分类精度影响较大的基因，而影响分类精度.采用的方法在其他分类模型和数据集上都有较好的效果.

为避免实验的偶发性，进行重复实验，针对平均分类精度和特征基因平均选择数量进行比较，并进行分类精度的方差分析，采用 SPSS 工具对平均分类精度进行方差分析，如表 3 所示.表 3 中 Aca/%表示平均分类精度，Avq/N 表示平均选择的特征基因数量，Ava 表示分类精度的方差.从表 3 可以看出，本文所采用的特征选择方法在选择不同特征基因数量的分类精度时相对稳定.

除此之外，为更好地验证本文所提 PCA-IG 算法的性能，实验采用 3 种常用分类算法进行分类实验验证，分类算法分别是 SGD、C4.5 和 RandomForest.其对比对象是直接将原始数据作为特征基因和经 PCA-IG 算法所得到的特征基因在 SGD、C4.5 和 RandomForest 3 种分类模型中实验的结果，数据处理前后在不同分类算法上的分类精度对比结果如表 4 所示.

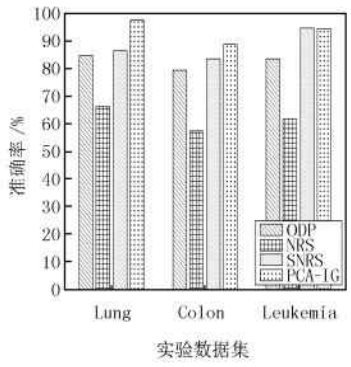

图 2 C4.5 在不同数据集上的分类性能

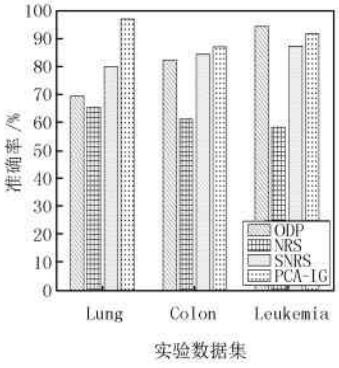

图 3 Naive Bayesian 在不同数据集上的分类性能

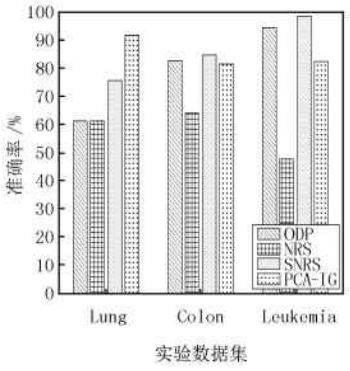

图 4 LibSVM 在不同数据集上的分类性能

表 3 3 种分类算法的平均分类精度、平均选择数量和平均分类精度方差

分类算法	Lung			Colon			Leukemia		
	Aca/%	Acq/N	Ava	Aca/%	Acq/N	Ava	Aca/%	Acq/N	Ava
C4.5	97.24	30	0.20	88.72	17	0.01	93.76	7	0.42
Naive Bayesian	97.24	30	0.08	86.78	17	0.53	91.40	7	0.34
LibSVM	91.63	30	0.00	86.78	17	0.53	81.86	7	0.28

表 4 数据处理前后在不同分类算法上的分类精度对比表

分类算法	Lung		Colon		Leukemia	
	原始数据	PCA-IG	原始数据	PCA-IG	原始数据	PCA-IG
SGD	98.03	96.06	82.26	87.10	94.44	94.44
C4.5	89.01	97.54	75.80	88.71	79.16	93.06
RandomForest	97.04	98.03	82.26	90.32	86.11	94.44

注:表 4 中 PCA-IG 为本文提出的基于 PCA 和信息增益的特征基因选择方法,数据表示分类精度(%).

从表 4 可看出,在 Lung 数据集中,在 C4.5 和 RandomForest 上的分类精度均高于将原始数据直接作为特征基因的分类精度,在 C4.5 上比将原始数据直接作为特征基因的精度高 8.53%,而在 SGD 上所得分类精度略低于将原始数据直接作为特征基因的分类精度 1.97%;在 Colon 数据集中,SGD、C4.5 和 RandomForest 3 种模型的分类精度,均高于将原始数据直接作为特征基因在 Weka 中的分类精度,其中在 C4.5 上所得分类精度高于将原始数据直接作为特征基因的分类精度 12.91%;在 Leukemia 数据集中,在 C4.5 和 RandomForest 上的分类精度均高于将原始数据直接作为特征基因的分类精度,其中在 C4.5 上所得分类精度高于将原始数据直接作为特征基因的分类精度 13.9%,而在 SGD 上,两种分类精度相同.实验结果表明采用 PCA 和 IG 进行无关信息过滤,可有效剔除冗余信息,选择出关联度高且低冗余度的特征基因,从而尽可能多地保留有效基因,采用不同算法所得的分类精度对比可证明算法的有效性.

表 5 选取 ODP、PCA 和 IG 3 种传统的特征基因选择方法及 NRS、SNRS[24] 和 SNRRF[14] 改进的特征基因选择方法与本文提出的 PCA-IG 特征基因选择方法的分类精度及选择特征基因数目作对比.根据表 7 可看出,在 Lung 基因数据集上,所提出的 PCA-IG 特征基因选择方法最优分类精度为 98.03%、特征基因数目为 20 个;在 Colon 基因数据集上,文中采用的 PCA-IG 特征基因选择方法最优分类精度为 90.32%、特征基因数目为 17 个;在 Leukemia 基因数据集上,本文方法的最优分类精度为 94.44%、特征基因数目为 6 个.其中在肺癌基因数据集上,PCA-IG 特征基因选择方法的分类精度比 PCA 算法高 26.38%且比 SNRS 算法高 12.59%;在结肠癌基因数据集上,PCA-IG 特征基因选择方法比单一使用 PCA 处理的特征基因数据后的分类精度高 28.79%,比 SNRRF 分类精度高 2.84%;在白血病肿瘤基因数据集上,PCA-IG 特征基因选择方法高于单一使用 PCA 处理的特征基因数据后的分类精度高 38.27%,但比 SNRRF 的分类精度低 0.33%.PCA-

IG 算法在提高分类精度的同时，经 PCA-IG 基因选择方法选择的基因特征子集数目明显减少，有效地剔除冗余基因.以 Leukemia 基因数据集和 Colon 基因数据集为例，经过多次实验筛选出现频率较高的部分基因及基因描述，如表 6～7.文献[25]验证解除抗 CD33 在治疗成人和儿童急性骨髓性白血病中有着很大的潜力.文献[26]证明在骨髓性白血病中，Zyxin 抑制降低抗凋亡蛋白 BCL2 和 BCL-XL 的表达.文献[27]提供了钙粘蛋白聚糖如何构成肿瘤生物和潜在治疗靶细胞的观点.文献[28]得出结论全长 WDSV 克隆或 orf 基因的表达抑制宿主鱼和人类肿瘤细胞生长，而证实 orf 基因相对于人类肿瘤细胞生长的重要性.

最后，为进一步验证本文提出的 PCA-IG 基因选择方法的分类性能，本文所采用的特征基因选择方法的分类精度与其他特征基因选择方法的分类精度进行最优分类精度对比，对比结果如表 5 所示.

表 5　本文特征基因选择方法和其他特征基因选择方法最优分类精度对比表

数据集	ODP	PCA	IG	NRS	SNRS	SNRRF	PCA-IG
Lung	12 600/91.62	202/71.65	146/97.23	128/88.36	6/85.44	10/89.89	20/98.03
Colon	2 000/64.51	61/61.53	26/89.25	51/73.24	6/82.26	72/87.48	17/90.32
Leukemia	7 129/94.44	71/56.27	69/94.03	37/67.38	4/97.36	26/94.77	6/94.44

注：表 5 中斜杠左侧数据是各特征基因选择方法进行分类的特征基因数(个)，右侧数据是各特征基因选择方法的最优分类精度(%).

表 6　本文算法在 Leukemia 基因数据集中选择的部分特征基因

序号	基因 ID	基因描述
1	M23197	CD33 antigen(differentiation antigen)
2	D84294	TPRD
3	L5148	Protein tyrosine kinase related mRNA sequence
4	X95735	Zyxin
5	U49395	Ionotropic ATP receptor P2X5a mRNA

表 7　本文算法在 Colon 基因数据集中选择的部分特征基因

序号	基因 ID	基因描述
1	R62549	PUTATIVE SERINE/THREONINE-PROTEIN KINASE B0464.5 IN CHROMOSOME III
2	H6524	GELSOLIN PRECURSOR,PLASMA(HUMAN)
3	D13641	Human mRNA for ORF ,complete cds
4	R44770	METABOTROPIC GLUTAMATE RECEPTOR 2 PRECURSOR
5	X63629	H.sapiens mRNA for p cadherin

4　结　论

本文结合 PCA 和信息增益方法，提出了一种基于 PCA 和信息增益的肿瘤特征基因选择方法(PCA-IG 算法).该方法首先采用 PCA 算法可有效地对原始数据进行降维，获得降维后的预选基因子集；然后通过信息增益获取较优的特征基因子集.实验结果表明，所提出的 PCA-IG 算法可提高基因表达谱的分类精度，进而有效地筛选出致病基因.

参　考　文　献

[1]　于化龙，顾国昌，赵靖，等.基于 DNA 微阵列数据的癌症分类问题研究进展[J].计算机科学，2010，37(10)：16-32.

[2]　Golub T R，Slonim D K，Tamayop，et al.Class discovery and class prediction by gene expression monitoring[J].Science，1999，286(2)：531-537.

[3]　Krishnapuram B，Carin L，Hartemink A.Gene expression analysis：Joint feature selection and classifier design[M].Massachusetts：MIT Press，2004：299-318.

[4]　Chen W，Zheng R，Baade P D，et al.Cancer statistics in China，2015[J].CA Cancer J Clin，2016，66(2)：115-132.

[5] 汪荆琪,徐林莉.一种基于多视图数据的半监督特征选择和聚类方法[J].数据采集与处理,2015,30(1):106-116.
[6] Xing E P,Jordan M I,Karp R M.Feature selection for high-dimensional genomic microarraydata[C].Proceedings of the 18th international conference on Machine Learing,Williamstown,2001.
[7] Callow M J,Dudoit S,Gong E L,et al.Microarray expression profiling identifies genes with altered expression in HDL-deficient mice[J].Genome Res,2000,10(12):2022-2029.
[8] 胡洋,李波.基于 Fisher 准则和多类相关矩阵分析的肿瘤基因特征选择方法[J].计算机应用与软件,2016,33(7):76-79.
[9] 陈涛,洪增林,邓方安.基于优化的邻域粗糙集的混合基因选择算法[J].计算机科学,2014,41(10):291-294.
[10] 徐天贺,马媛媛,徐久成.一种基于邻域互信息最大化和粒子群优化的特征基因选择方法[J].小型微型计算机系统,2016,8(37):1775-1779.
[11] 谢娟英,高红超.基于统计相关性与 K-means 的区分基因子集选择算法[J].软件学报,2014,25(9):2050-2075.
[12] 魏莎莎,陆慧娟,安春霖,等.一种基于互信息最大化的模型无关基因选择方法[J].计算机科学,2014,41(9):224,243-247.
[13] 吴辰文,王伟,李长生,等.一种结合随机森林和邻域粗糙集的特征选择方法[J].小型微型计算机系统,2017,6(38):1358-1362.
[14] 徐久成,冯森,穆辉宇.基于信噪比与随机森林的肿瘤特征基因选择[J].河南师范大学学报(自然科学版),2017,45(2):87-92.
[15] 沈宁敏,李静,周培云,等.一种基于稀疏主成分的基因表达数据特征提取方法[J].计算机科学,2015,42(6A):453-458.
[16] 张玉春,郝平波,王明宇,等.结肠癌基因表达谱的分类检测问题研究[J].计算机工程与应用,2011,47(17):231,244-248.
[17] 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.
[18] 刘庆和,梁正友.一种基于信息增益的特征优化选择方法[J].计算机工程与应用,2011,47(12):130-136.
[19] 张小康,帅建梅,史林.基于加权信息增益的恶意代码检测方法[J].计算机工程,2010,36(6):149-151.
[20] 芮兰兰,张洁,郭少勇,熊翱.基于样本加权的基因特征选择模型[J].北京邮电大学学报,2016,39(s1):72-75.
[21] 王洪喜,彭宏.一种基于主成分分析的异常点挖掘方法[J].计算机科学,2007,34(10):192-194.
[22] 阮越,陈汉武,刘志昊,等.量子主成分分析算法[J].计算机学报,2014,37(3):666-676.
[23] 李航.统计学习方法[M].北京:清华大学出版社,2012.
[24] 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,2015,5(30):973-981.
[25] Laing A A,Harrison C J,Gibson B E S,et al.Unlocking the potential of anti-CD33 therapy in adult and childhood acute myeloid leukemia [J].Experimental Hematology,2017,54:40-50.
[26] Bernusso V A,Machado-Neto J A.Pericole FV Imatinib restores VASP activity and its interaction with Zyxin in BCR-ABL leukemic cells [J].Biochimica Et Biophysica Acta-molecular Cell Research,2015,1853(2):388-395.
[27] Carvalho S,Reis C A,Pinho S S.Cadherins Glycans in Cancer:Sweet Players in a Bitter Process[J].Trends Cancer,2016,2(9):519-531.
[28] Xu K,Zhang T T,Wang L,et al.Walleye dermal sarcoma virus:expression of a full-length clone or the rv-cyclin (orf a) gene is cytopathic to the host and human tumor cells[J].Molecular Biology Reports,2013,40(2):1451-1461.

Tumor feature gene selection method based on PCA and information gain

Xu Jiucheng, Huang Fangzhou, Mu Huiyu, Wang Yun, Xu Zhanwei

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

Abstract: Aiming at the low classification accuracy of tumor genetic data with the characterstic of high dimensional and unrelated genes, a tumor feature gene selection method based on PCA and information gain is proposed. Firstly, the PCA algorithm is used to eliminate miscellaneous genes and select the preselected feature gene subset in this method. Then, the information gain algorithm is used to optimize the subset of the preselected feature gene subset, and the feature gene subset is obtained. Finally, different sorting algorithms are used to simulate the feature gene subset. The experimental results show that the method proposed in this paper improves the classification accuracy of gene expression profile, thus indicating that the pathogenic gene is effectively selected.

Keywords: gene classification; PCA; information gain; feature selection

[责任编校 陈留院]