文章编号:1000－2367(2019)02－0021－08 **DOI**:10.16366/j.cnki.1000－2367.2019.02.004

基于 K-S 检验和邻域粗糙集的特征选择方法

刘艳,程璐,孙林

(河南师范大学 计算机与信息工程学院,河南 新乡 453007)

摘 要:传统的肿瘤基因选择算法挑选出的特征基因中存在大量噪声基因和冗余基因,从而对基因算法的准确性和分类精度产生影响.针对这一问题,将 K-S 检验与邻域粗糙集融合成为一种新的特征选择方法.首先,采用累积分布函数计算正负类样本的累积函数值和 K-S 检验统计量,对照显著性水平下的样本统计量,从而去除冗余基因和噪声基因;然后,使用邻域粗糙集进行约简,对比条件属性重要度得出最优约简结果;最后,对比 K-S 检验和两种基于 K-S 检验的特征选择方法得到的冗余度和分类精度,通过实验验证这种方法不仅能准确挑选出具有显著区分能力的肿瘤基因,且效率高具有可行性.

关键词:K-S 检验;邻域粗糙集;特征选择

中图分类号:TP181 **文献标志码**:A

近年来,DNA 芯片技术的迅速成熟以及基因表达谱数据(GEP)的成倍增长为人类身体状况的分析和疾病的诊治提供了有效帮助[1].然而,在微阵列数据的大量基因中,只有一小部分的重要基因可作为生物标志物追踪疾病[2].到目前为止,在基因序列、基因芯片以及利用基因微列阵技术进行肿瘤病情判断和诊治等方面已经取得长足发展,很多学者专注于不同的方面进行了深入研究和学习,并将其运用到实际生活当中,取得了不错的成果.但是不可否认,目前的技术在基因分类和基因选择方面仍然存在很多问题.基因微阵列技术的出现对于癌症基因的病情判断和早期控制产生了重要影响,不仅有利于快速识别出特征属性,同时还有助于提高判断病情的准确性,使病人获得更好的临床诊断,尤其是在诊断肿瘤病情方面.

通过近年来针对基因数据的实验研究可以得知,基因数据分析主要涵盖了 4 个步骤,即基因数据获取、基因数据预处理、基因选择、分类模型建立与评估[3].其中,基因特征选择在对肿瘤基因数据进行筛选的过程中十分关键,也就是运用合理的算法从大量的基因表达数据中选取部分典型的、有代表性的、对疾病的预测和诊断有比较强的鉴别能力的特征基因.将特征选择的类别分为 3 大类:过滤式(Filter) 、缠绕式(Wrapper) 、嵌入式(Embedded).其中,前两种特征选择方法在日常研究中是使用最为频繁和便利的,而这两种方法的具体区别在于学习过程是否独立[4].首先,最常被使用的过滤式方法在基因选择时,不仅节约了大量时间,提高了工作效率,而且能够根据单个判别标准选取相对重要的,即区分能力更强的特征基因作为基因子集,例如基于信噪比、基于 t-统计等过滤法[4].其次,缠绕式方法是参照预设的学习算法的特性来评估和挑选出特征结果,其结果可以比前者更加优化,但是也存在计算量较大,通用性较差,在执行算法的过程中易陷入部分最优的情况[5].将过滤式和缠绕式相结合的混合特征选择方法,在选择过程中所需时间更短,速度更快,但是在剔除冗余基因方面的效果没有那么好[5].而嵌入式方法不同于过滤式和缠绕式,该方法将基因选择嵌入到具体的学习算法里,这样就可以在一边进行学习模型的训练,一边完成对特征基因的筛选[6].

收稿日期:2018-06-15;**修回日期**:2018-12-19.

基金项目:国家自然科学基金(61772176);中国博士后科学基金项目(2016M602247);河南省科技创新人才项目(184100510003);河南省科技攻关项目(182102210362);河南省高校青年骨干教师培养计划项目(2017GGJS041).

作者简介:刘艳(1983－),女,河南郑州人,河南师范大学实验师,研究方向为数据挖掘、图像处理, E-mail:liu_yan122@sina.com.

通信作者:孙林,E-mail:sunlin@htu.edu.cn.

Kolmogorov-Smirnov(K-S 检验)是用于对比两类样本是否属于同一分布的常用非参数统计方法[7],在识别两类不同样本的分布形状差异时表现得十分灵敏,在卵巢癌基因数据的分析[8]、情感识别等领域取得一定发展和突破[9].谢娟英等[3]对 K-S 检验进行扩展应用,将其与 mRMR 结合在一起,应用于特殊基因类别的选择.胡秋锋[10]则在 K-S 检验的基础之上,充分学习了多种基因选择算法的理论,然后通过实验验证了游程检验与 K-S 检验在基因选择中的实际应用.而近些年来,关于粗糙集的研究已经成为一种趋势,愈来愈多的研究者开始将这一理论运用在不同的领域当中,这一理论自提出以后就在不同的应用方面取得系列突出成就,其中以胡清华为主要代表.它已经成为一种新型的数学工具,能够应用于描绘出那些不确定的信息分类.而且有大量英文文献指出,这种工具不需要任何其他附加条件就能够找到一个范围最小且与全部属性具有相同区分能力的属性子集,这就是通过属性约简来最大限度地提高运算速度[11-15].胡清华等[16]基于该理论提出邻域粗糙集模型,这一创新点在敏感特征的选择方面取得了较好的效果,其中邻域半径的大小与相关参数有关,即阈值的不同设置,直接影响着最终的分类精度和提取的特征基因数.Hu 等[17]基于粗糙集理论将其构建在不同的分布式环境下,对此类新的分布式决策信息系统赋予了全新的定义,并参照这种定义创新性地提出了不同于传统方式的分布式决策信息系统属性约简算法.因此,结合国内外发展现状,邻域粗糙集在各个领域的发展都取得了不同程度的进展,结合基因选择的发展现状,该理论在肿瘤基因特征选择的实际利用上确实十分有前景.

本文根据已有的大量有关邻域粗糙集理论的文献资料,结合 K-S 检验,将 K-S 检验和邻域粗糙集同时运用到基因选择当中.首先,为了使筛选过后的特征基因具有明显的区分能力,采用 K-S 检验剔除原始基因数据集中的大量冗余和噪声基因,同时种群基因的多样性增强;然后,预选择子集经过约简和对比条件属性的重要度参数,可以将基因间存在的某些相关性保存下来,这改良了传统的独立非参数检验未能充分考虑基因间冗余的情况,挑选出最优的基因子集;最后,通过对冗余度的定义,结合本文的基因数据集,对基因集合的冗余度进行计算;两种方法通过对比实验证明了基于 K-S 检验和邻域粗糙集的特征选择方法切合实际且有效.

1 相关知识

1.1 K-S 检验

假定原始集合中涵盖了两类不同的样本,分别为正类(A)和负类(B)这两组独立的类别.基因数据集中的总样本数目设为 n,以原始基因数据集中的某个基因 X 为例,设 $X_1,X_2,\cdots,X_n$ 是从基因 X 中取出的,那么可以得知该基因 X 的观测值就是 $x_{(1)},x_{(2)},\cdots,x_{(n)}$.参照特征值对其降序排列可以获取其次序观测值 $x_{(1)}\leq x_{(2)}\leq\cdots\leq x_{(n)}$,由此可以得出该基因的累积分布函数[18],公式如下:

$$F(x)=\begin{cases}0, & x<x_{(1)},\\ \dfrac{k}{n}, & x(k)\leq x\leq x(k+1),k=1,2,\cdots,n-1,\\ n, & x\geq x(n).\end{cases}\tag{1}$$

K-S 检验统计量 T 是每个正类样本的累积分布函数 $F_{A(x)}$ 与每个负类样本的累积分布函数 $F_{B(x)}$ 差值的绝对值的最大值,公式如下:

$$T=\max_x|F_A(x)-F_B(x)|.\tag{2}$$

在显著性水平为 α 的条件下,分别把样本分布的统计量 T 和假设的分布统计量 T_{crit} 进行比较.根据 K-S 检验理论,T_{crit} 为显著性水平 α 下样本统计量的临界值,计算过程中,如果某个样本统计量 T 不小于 T_{crit},那么就可以得出结论,即该基因的正类与负类样本的实际分布形状有明显的不同;而如果某个样本统计量的值 T 比 T_{crit} 的值更小,也可以得出结论,即该基因的正类与负类样本的实际分布形状并不存在明显的不同.

1.2 邻域粗糙集

本文参照文献[16, 19]给出与邻域粗糙集相关的具体概念和性质.

定义 1 给定大小为 N 维的实数空间 U,且设置 $U=\{x_1,x_2,\cdots,x_n\}$ 为非空实数集合,$\Delta:\mathbf{R}^N\times\mathbf{R}^N\to\mathbf{R}$,

则称 Δ 属于 $\mathbf{R}^N$ 上的一个度量值,如果 Δ 满足:

(1)$\Delta(x_1,x_2)\geq 0$,$\Delta(x_1,x_2)=0$,当且仅当 $x_1=x_2$,$\forall x_1,x_2\in\mathbf{R}^N$;

(2)$\Delta(x_1,x_2)=\Delta(x_2,x_1)$,$\forall x_1,x_2\in\mathbf{R}^N$;

(3)$\Delta(x_1,x_3)\leq\Delta(x_1,x_2)+\Delta(x_2,x_3)$,$\forall x_1,x_2,x_3\in\mathbf{R}^N$.

称 $\langle U,\Delta\rangle$ 为度量空间,一般情况下实属空间上的距离可用表示如下:

$$\Delta(x_i,x_j)=\left[\sum_{k=1}^{N}|x_{ik}-x_{jk}|^P\right]^{1/P}. \tag{3}$$

若 $P=1$ 时,Δ 表示为 City-Block 距离;若 $P=2$ 时,Δ 表示为欧式距离;若 $P=\infty$ 时,Δ 表示为 Dominance 距离.

定义 2　提前设置一个决策信息系统 $\langle U,A,D\rangle$,论域 $U=\{x_1,x_2,\cdots,x_n\}$,属性集合 A,且 $B\subseteq A$,则对于任意 $x_i\in U$,其 ε- 邻域 $\varepsilon_B(x_i)$ 被定义为

$$\varepsilon_B(x_i)=\{x_j\mid x_j\in U,\Delta_B(x_i,x_j)\leq\varepsilon\}. \tag{4}$$

(4)式中的 ε 属于邻域阈值,且 $\varepsilon\geq 0$,$\Delta_B(x_i,x_j)$ 表现为一种函数关系,反映样本 x_i 和样本 x_j 之间的距离关系,也可以将这种函数视为两者之间的相似程度的反映.而在整个论域 U 里面,将任何一个到 x_i 的距离小于邻域阈值 ε 的样本构成的这个集合,称为 x_i 的邻域.因此,一个邻域样本的集合通常要受到距离函数 Δ,属性集合 B,以及阈值 ε 的影响.另外邻域样本空间的大小往往会随着邻域半径的变化而变化,即邻域半径变大时,空间中样本数将相应地增多,同时样本之间存在的相似度就会降低.

性质 1　如果邻域的阈值 $\varepsilon_1\leq\varepsilon_2$,对于任意的 $x_i\in U$,则有 $\varepsilon_1(x_i)\subseteq\varepsilon_2(x_i)$.

性质 2　设置邻域阈值 ε,若 $B_1\subseteq B_2\subseteq C$,对于任意的 $x_i\in U$,则有 $\varepsilon B_1(x_i)\subseteq\varepsilon B_2(x_i)$.

定义 3　给定 $\langle U,A,D\rangle$ 为决策信息,$B\subseteq A$,决策 D 将论域 U 分别划为 n 个等价类,记为:$U/D=\{x_1,x_2,\cdots,x_n\}$. 在邻域空间中,它的下近似和上近似分别定义为

$$\underline{N}_B^Z(X_i)=\{x_i\in U\mid\varepsilon(x_i)\subseteq X_i\}, \tag{5}$$

$$\overline{N}_B^Z(X_i)=\{x_i\in U\mid\varepsilon(x_i)\cap X_i\neq\varnothing\}. \tag{6}$$

下近似 $\underline{N}_B^Z(X_i)$ 是指依照已经形成的系统知识判断,样本空间中完全隶属 X_i 的样本所构成集合的最大值.同样上近似 $\overline{N}_B^Z(X_i)$ 是指根据现有知识判断,样本空间中由那些属于或者不完全属于 X_i 的样本所组成集合的最小值.

定义 4　设定 $\langle U,A,D\rangle$ 为决策信息系统,并且决策 D 把论域 U 划分成了 n 个等价类:$X_1,X_2,\cdots,X_n$,那么对任意的 $B\subseteq A$,决策属性 D 关于 B 的下近似与上近似可定义为:

$$\underline{N}_B^Z(D)=\bigcup_{i=1}^{n}\underline{N}_B^Z(X_i), \tag{7}$$

$$\overline{N}_B^Z(D)=\bigcup_{i=1}^{n}\overline{N}_B^Z(X_i), \tag{8}$$

这里 $\varepsilon(x_i)$ 是通过属性集 B 和度量函数 Δ 共同产生的邻域粒子,而且决策 D 的下近似也被叫作决策正域,可以表示为:

$$NPOS_B^Z=\bigcup_{X_i\in U/D}\underline{N}_B^Z(X_i). \tag{9}$$

同样,决策的边界定义为:

$$BNR_B^Z=\overline{N}_B^Z(D)-\underline{N}_B^Z(D), \tag{10}$$

决策边界 $BNR_B^Z(D)$ 是指一个集合,通过参照已经存在的系统知识,可能辨别出其值是否包含在 X 中,但不能全面肯定其值是否一定存在于属于 X 的样本构成的集合.

容易验证满足:

(1)$\forall D\subseteq U$,有 $\underline{N}_B^Z(D)\subseteq X\subseteq\overline{N}_B^Z(D)$;

(2)$\forall B_1\subseteq B_2$,有 $\underline{N}_B^Z(D)\subseteq\overline{N}_B^Z(D)$;

(3)$\forall B_1\subseteq B_2$,存在 $NPOS_{B_1}^Z(D)\subseteq NPOS_{B_2}^Z(D)$.

在决策信息系统中,上、下近似是粗糙集理论中最基本的概念,并在邻域粗糙集领域得到了有效发展,通

常用上、下近似来刻画决策属性的等价类.同时,众多学者参照上、下近似概念构造出基于邻域粗糙集的正域、边界等概念.

定义 5 给定一个决策信息系统 $\langle U,A,D\rangle$,决策属性 D 相对于条件属性 B 而言得出的依赖度可以用公式表示如下:

$$\eta_B^Z(D)=\frac{NPOS_B^Z(D)}{|U|}. \tag{11}$$

定义 6 给定一个决策信息系统 $\langle U,A,D\rangle$,Δ 是属于 U 上的度量,对于任何的 $B\subseteq A$,如果满足在 ε 粒度下的条件,则属性 B 满足:

(1)$\eta_B^Z(D)=\eta_A^Z(D)$;

(2)$\forall a\in B:\eta_{B-a}^Z(D)<\eta_B^Z(D)$,

则称 B 是 ε 粒度下的相对约简.

2 基于 K-S 检验和邻域粗糙集的特征选择方法

针对前面提到过的两种传统特征选择方法在基因选择中的不同作用,本文将 K-S 检验和邻域粗糙集理论联合起来,融合成为有关基因选择的算法(Feature Selection Based on K-S and Neighborhood Rough Sets, KSNRSFS),采用 K-S 检验算法降低基因维数,缩小搜索范围,利用邻域粗糙集进行特征基因的属性约简,从而挑选出最优特征基因.详细步骤见算法 1.

算法 1 一种基于 K-S 检验和邻域粗糙集理论的特征选择算法

输入 基因数据集 $X=\{A_1,A_2,\cdots,A_p,B_1,B_2,\cdots,B_q\}$,显著性水平 α,邻域决策系统 $\langle U,A,D\rangle$,邻域半径 δ 的参数值 λ 以及重要度下限参数 efc_ctrl;

输出 特征基因集合 feature-elect.

步骤 1 基因样本分为正类 A、负类 B 两种,将基因 X 的特征值 y 代入到累积分布函数公式(1),从而计算出正、负类基因样本的累积分布函数值 $F_A(x)$ 和 $B_B(x)$;

步骤 2 然后由公式(2)得到基因 X 的 K-S 检验统计量 T 的值,然后将其与 α 相符的临界值 T_{crit} 依次对照,若计算得出统计量 T 的值不小于 T_{crit},那么可将这个基因视为具有显著区分能力的基因,并且挑选出此类样本组合成为预选择基因集合 red;

步骤 3 对基因子集 red 进行归一化处理,$std\_\text{red}=std(\text{red})$;

步骤 4 归一化后整合数据结果,按照算法要求将条件属性放置在最靠前的一列,将决策属性放在最靠后的一列;

步骤 5 计算邻域半径 $\delta=std\_\text{red}(\text{red})./\lambda$;

步骤 6 对 $A\text{-}red$ 中每个属性 a_i 计算正域 $Pos_{\text{red}+a_i}^{smp}(D)$;

步骤 7 寻找最大正域 $Pos_k(D)$,从而得到条件属性重要度的结果;

步骤 8 若重要度大于设定的下限 efc_ctrl,则输出约简结果 red;若小于,返回步骤 6.

3 实验分析

3.1 实验数据

本文实验的数据集分别为:前列腺癌(Prostate)、白血病(Leukemia)和肺癌(Lung),数据集可从 UCI 和 GEMS 网站下载,关于数据集的详细信息见表 1.具体的硬件环境如下:计算机系统是 Windows10,操作系统属于 64 位、内存大小是 8 GB、处理器信息为 Intel(R) Core(TM) i7-4710HQ CPU @2.50 GHz.本文所有实验都是在相关软件(如 MatlabR2016b 和 weka3.9.0)中实现的.

3.2 实验结果与分析

针对两种与 K-S 检验相关的特征选择方法,本文以 lung 原始基因数据集为例,分别完成了基于 K-S 检验的 Relief 特征选择方法(Feature Selection Based on K-S test and Relief , KSReliefFS)[20] 和 KSNRSFS

的实验,然后对照通过两种实验方法得到的冗余度的值,详细实验过程如下.

(1)基于 K-S 检验的 Relief 特征选择算方法

实验过程中利用 MATLAB 将 K-S 检验的过程展现出来:由图像观察可知垂直方向上的基因的正类样本与负类样本间的差值,就是基因的显著性差异.以 Lung 原始基因数据集为例,共由 31 个正类样本和 39 个负类样本组成,而每个基因样本均有 1 626 个特征值.用 positive 表示正常样本的累积分布概率,用 negative 表示肿瘤样本的累积分布概率,其中横轴的 x 表示基因值,纵轴的 $F(x)$ 表示基因值 x 对应的累积分布概率,而同一基因值的正、负类样本的累积分布概率在垂直方向上的最大差值的绝对值,即为该基因的属性 X 的 K-S 检验的统计量 T.对比统计量 T 与显著性水平下的某个样本统计量的临界值 T_{crit} 即可确认该基因值是否能够帮助识别显著性差异,从而进行有效的基因筛选.如图 1 和图 2 所示,对照第 8 个基因和第 10 个基因的累积分布概率差值和最终的统计量 T,得知 K-S 检验能够灵敏地察觉到不同基因样本分布形状的差异,举例来说,第 10 个基因样本分布形状的差异显著大于第 8 个,因此,可以认为前者更有助于在进行基因筛选时剔除无关基因.结果显示从 Lung 原始基因数据集共 70 个样本中筛选出了区分能力较为明显的共 1 280 个基因样本,缩小了下一步进行筛选时要搜索的范围.

表 1　实验数据集描述

Tab.1　Description of experimental data sets

基因数据集	特征总数	样本数(正/负类)	类别
Prostate	12 600	54(23+31)	2
Leukemia	1 869	63(25+38)	2
Lung	1 626	70(31+39)	2

然后将初次筛选过后的特征基因放入一个新的集合中,用 Relief 算法进行特征选择,由表 2 可以得知,在基因权重阈值取不同值时,KSReliefFS 得到的最终 lung 特征值个数有所不同,而经过 weka 测试得到的 lung 基因表达谱数据集上的分类精度却没有发生变化,基本保持稳定,且分类精度高达 98.571 4%.针对 lung 基因数据集实验时,基因权重阈值分别设置为 4 000、5 000、6 000、7 000、8 000、9 000、10 000,这里利用了 KSReliefFS 来完成对原始基因集合的一系列特征筛选,并实时标记下挑选出的特征结果,然后采用了 weka 软件内附带的 SVM 分类器实现对最终结果的准确测验,分别记录下其分类精度数值[21].最后,为了探究基因权重阈值对最终特征基因个数的影响,本文通过多次试验,对基因的权

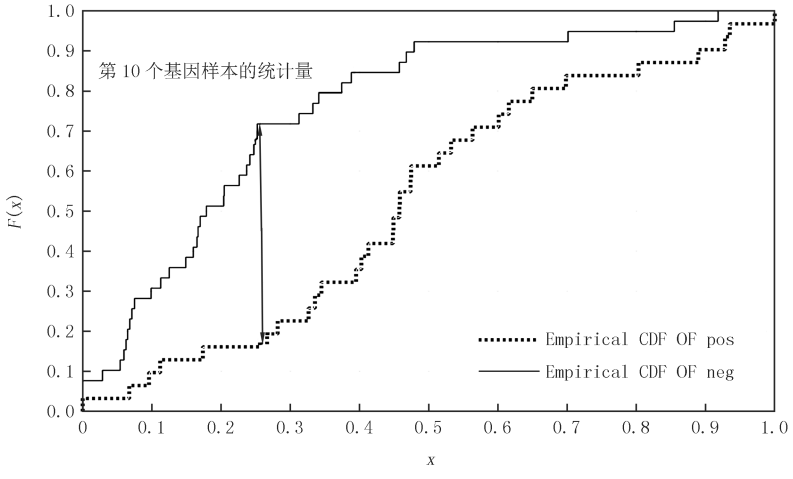

图 1　第 10 个基因样本的统计量

Fig.1　Statistics of the 10th gene sample

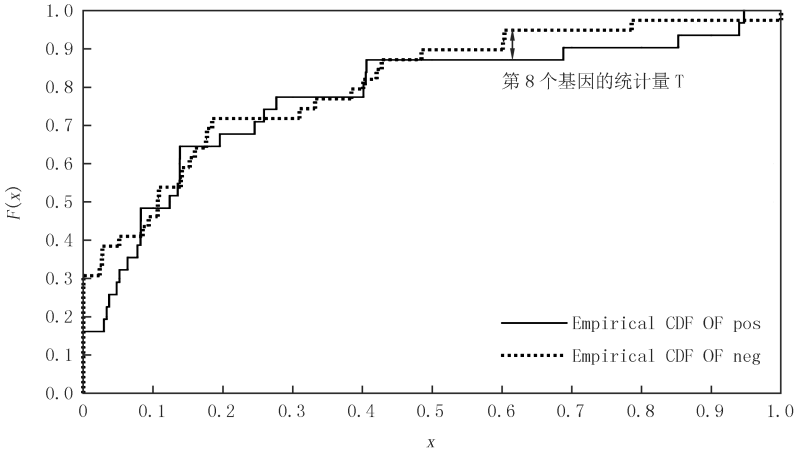

图 2　第 8 个基因样本的统计量

Fig.2　Statistics of the 8th gene sample

重阈值取不同的值,然后依次测验和记录其分类精度,从而方便进行对比.

而由图 3 和图 4 可知,采用了 KSReliefFS 方法之后,lung 原始基因数据集中的冗余特征呈现了逐渐降低的趋势,在实验过程中,筛选出的特征基因个数随着基因权重阈值的增大而逐步地减少,但分类精度稳定在 98.571 4%,并未随阈值的变化而发生明显变化,也就说明了一个问题,即 KSReliefFS 对基因权重阈值的取值十分不敏感,即使基因权重阈值有所差别,该算法也可以对分类后的基因数据集采取特征选择方法,然后通过 Weka 软件测试出最终分类精度的结果,从而评价该算法的分类性能.因此,通过本文实验,可以得出如下结论:即 KSReliefFS 具有良好稳定性和可行性.

(2)基于 K-S 检验和邻域粗糙集的特征选择方法

实验过程中,邻域半径受到 λ 取值得的影响,而邻域半径最终的结果又会影响约简结果.以 lung 和 prostate 基因数据集为例,经过多次试验,将 λ 的值设置在不同区间内,研究 λ 的取值给实验结果带来的变化,具体变化见表 3 和表 4.根据统计结果得知,lung 基因数据集的 λ 取值在(0.1,1)之间时,特征值稳定在 1 到 3 个,结合重要度取值分析,暂定 lung 基因数据集中 λ 值为 0.3;而 prostate 基因数据集的 λ 取值在(0,2)区间时,特征值和重要度比较稳定.因此,可以得知不同的基因数据集针对不同的 λ 值,要根据具体情况多次进行实验才能最终确定.

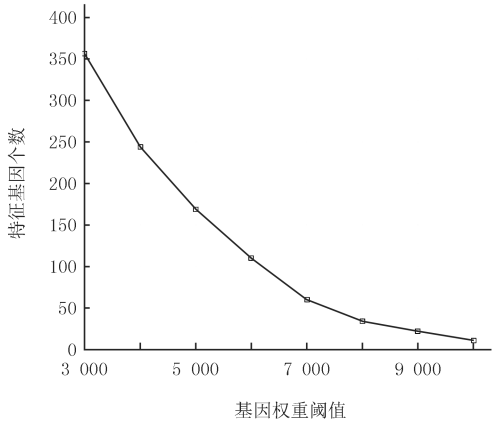

图 3 Lung 数据集基因权重阈值对应特征基因

Fig.3 Lung dataset gene weight threshold corresponding to the characteristic gene

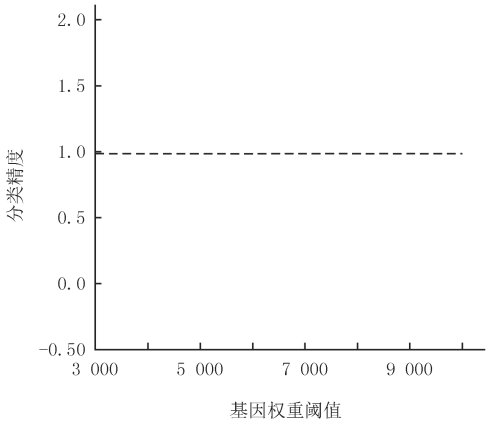

图 4 Lung 数据集基因权重阈值对应分类精度

Fig.4 Lung dataset gene weight threshold corresponding classification accuracy

表 2 基因权重阈值取值和对应的分类精度

Tab.2 Gene weight threshold value and corresponding classification accuracy

基因权重阈值	特征基因个数	分类精度/%
4 000	244	98.571 4
5 000	169	98.571 4
6 000	110	98.571 4
7 000	60	98.571 4
8 000	34	98.571 4
9 000	22	98.571 4
10 000	11	98.571 4

表 3 lung 基因集中 λ 取值、特征值和重要度

Tab.3 Value and eigenvalue and importance of λ in lung gene set

λ	特征值	重要度
0.1	398	0.329
0.2	1 233	0.671
0.3	1 233 22 48	0.871 0.957 1.000
0.4	1 233 22	0.943 1.000
0.5	1 233 22	0.943 1.000
0.6	1 233 22	0.943 1.000
0.7	1 233 22	0.943 1.000
0.8	1 233 4	0.971 1.000
0.9	1 233 4	0.971 1.000

(3)两种基于 K-S 检验的两种特征选择方法冗余度对比

为了对比两种特征选择方法的冗余程度,陈玉明等[22]定义冗余度来表示选择基因的精简程度,表示如下:

$$r=\frac{|R|}{|C|}. \tag{12}$$

用基因选择后的基因数目的绝对值除以原始基因数据的基因数目,那么 r 的值越接近 0,则表示冗余程度越低,精简的效果越明显.而对不同基因数据集使用不同特征选择方法的冗余度结果[21]见表 5.

实验结果显示,K-S test,KSNRSFS 以及 KSReliefFS 的结果相对比,其中得出的冗余度最高的是只进行 K- S 的基因子集,冗余度最低的是 KSNRSFS,以 prostate 基因数据集为例,3 种特征选择方法得到的冗余度分别为 98.7%、0.02%和 0.04%,且结合 KSNRSFS 在筛选时速度更快,最终结果更精确且具体,对比 3 种基因数据集的实验结果发现,KSNRSFS 具有可行性.结合表 6 可知,将 3 个数据集的实验结果分别放入 Weka 软件中的 4 种分类器 SVM,RandomTree,J48 和 PART 中进行精度测试,可以发现 KSNRSFS 所得精度普遍较高且在不同的分类器中得出的分类精度结果也比较稳定,由此证明本文提出的特征选择方法是有效的.

表 4　prostate 基因集中 λ 和特征值

Tab.4　λ and eigenvalues of the prostate gene set

Lammda	特征值
1.5	144　4 700　6 853
1.6	402　6 044　7 809
1.7	18　4 700　9 727
1.8	10　4 453　8 682
1.9	2　4 453　8 682
2.0	2　4 700　8 682

表 5　K-S test、KSNRSFS 与 KSReliefFS 冗余度结果对比

Tab.5　Comparison of redundancy results among K-S test, KSNRSFS and KSReliefFS

基因数据集	特征选择方法	特征值数目	冗余度/%
Prostate	K-S test	12 437	98.7
	KSNRSFS	3	0.02
	KSReliefFS	5	0.04
Leukemia	K-S test	1 409	75.4
	KSNRSFS	5	0.27
	KSReliefFS	33	1.77
Lung	K-S test	1 280	78.7
	KSNRSFS	3	0.18
	KSReliefFS	11	0.67

表 6　K-S,KSNRSFS 与 KSReliefFS 结果分类精度对比

Tab.6　Comparison of classification accuracy among K-S, KSNRSFS and KSReliefFS

基因数据集	特征选择方法	SVM	RandomTree	J48	PART
Prostate	K-S test	57.407 4%	68.518 5%	85.185 2%	85.185 2%
	KSNRSFS	81.431 5%	75.925 9%	85.185 2%	85.185 2%
	KSReliefFS	57.407 4%	88.888 9%	90.740 7%	90.740 7%
Leukemia	K-S test	75.000 0%	77.777 8%	77.777 8%	75.000 0%
	KSNRSFS	77.777 8%	73.611 1%	75.000 0%	75.000 0%
	KSReliefFS	58.730 2%	71.428 6%	69.841 3%	73.015 9%
Lung	K-S test	74.285 7%	82.857 1%	98.571 4%	98.571 4%
	KSNRSFS	92.857 1%	94.285 7%	98.571 4%	98.571 4%
	KSReliefFS	100.000 0%	92.587 1%	98.571 4%	98.571 4%

4　结束语

K- S 检验可以去除噪声基因和部分冗余基因,缩小基因选择时要搜索的范围,挑选出预选择基因子集,邻域粗糙集算法利用条件属性的重要度对比,删除无关的冗余特征,快速筛选出约简结果.首先,运用 K-S 检验中的累积分布函数对基因数据集进行降维处理,缩小搜索空间;然后,采用邻域粗糙集的前向贪心算法,计算每个条件属性的正域集合,并对比重要度,约简出最优特征集合;最后,对比 KSReliefFS 和 KSNRSFS,依照两种方法实验结果的优劣性可知,前者的冗余度以及时间复杂度都大于后者,从而验证了该算法的可行性,能够快速获得较少的特征基因.

参 考 文 献

[1] 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.

[2] Sun L,Zhang X Y,Xu J C,et al.A gene selection approach based on the fisher linear discriminant and the neighborhood rough set[J].Bioengineered,2018,9(1):144-151.

[3] 谢娟英,胡秋锋,董亚非.K-S 检验与 mRNR 相结合的基因选择算法[J].计算机应用研究,2016,33(4):1001-3695.

[4] Sun L,Xu J C,Wang W,et al.Locally linear embedding and neighborhood rough set-based gene selection for gene expression data classification[J].Genetics and Molecular Research,2016,15(3):15038990.

[5] 张新乐.基于邻域粗糙集的特征选择方法研究[D].新乡:河南师范大学,2018.

[6] 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,2016,37(8):1775-1779.

[7] Charmpi K,Ycart B.Weighted Kolmogorov Smirnov testing:an alternative for Gene Set Enrichment Analysis[J].Statistical Applications in Genetics and Molecular Biology,2015,14(3):279-293.

[8] Sun L,Zhang X Y,Qian Y H,et al.Joint neighborhood entropy-based gene selection method with fisher score for tumor classification[J].Applied Intelligence,2018.DOI:10.1007/s10489-018-1320-1.

[9] 张丽娟,李舟军.微阵列数据癌症分类问题中的基因选择[J].计算机研究与发展,2009,46(5):794-802.

[10] 胡秋锋.游程检验与 K-S 检验在基因选择中的应用研究[D].西安:陕西师范大学,2015.

[11] 孙林,潘俊方,张霄雨,等.一种基于邻域粗糙集的多标记专属特征选择方法[J].计算机科学,2018,45(1):173-178.

[12] 徐久成,黄方舟,穆辉宇,等.基于 PCA 和信息增益的肿瘤特征基因选择方法[J].河南师范大学学报(自然科学版),2018,46(2):104-110.

[13] 孙林,刘弱南,张霄雨,等.一种基于粗糙均方残基的模糊双聚类方法[J].河南师范大学学报(自然科学版),2017,45(5):93-100.

[14] 王思华,杨桐,段启凡,等.基于 DT 法和粗糙集理论的接地网安全性状态评定[J].电力系统保护与控制,2017,45(2):48-54.

[15] 王振浩,杜虹锦,李国庆,等.基于 t-分布邻域嵌入的同调机群无监督识别[J].电力系统保护与控制,2018,46(22):64-71.

[16] Hu Q H,Yu D R,Liu J F,et al.Neighborhood rough set based heterogeneous feature subset selection[J].Information Sciences,2008,178(18):3577-3594.

[17] Hu J,Pedrycz W,Wang G Y,et al.Rough sets in distributed decision information systems[J].Knowledge-Based Systems,2016,94:13-22.

[18] Huang S G,Yeo A A,Li S.Modification of Kolmogorov-Smirnov test for DNA content data analysis through distribution alignment.Assay and Drug Development Technologies,2007,5(5):663-672.

[19] 陈文刚.基于邻域粗糙集属性约简算法的研究[D].辽宁:渤海大学,2017.

[20] 程璐,李欣,王薇,等.基于 K-S 检验的 Relief 特征基因选择方法[J].无线互联科技,2017,113(13):103-104.

[21] DeRisi J L,Vishwanath R I,Patrick O.Brown.Exploring the metabolic and genetic control of gene expression on a genomic scale[J].Science,1997,278(5338):680-686.

[22] Chen Y M,Zhang Z J,Zheng J Z,et al.Gene selection for tumor classification using neighborhood rough sets and entropy measures[J].Journal of Biomedical Informatics,2017,67:59-68.

Feature selection method based on K-S test and neighborhood rough sets

Liu Yan,Cheng Lu,Sun Lin

(College of Computer and Information Engineering,Henan Normal University,Xinxiang 453007,China)

Abstract:Traditional tumor gene selection algorithms usually remain many noisy and redundant genes in selected feature values,which affect the gene algorithm accuracy and the classification precision.Aiming at solving this the problem,we propose to combine the K-S test with neighborhood rough sets theory.Firstly,the cumulative distribution function is used to calculate the positive and negative cumulative distribution values and the K-S test statistic,and the sample statistics under the significance level are compared to remove those redundant and noisy genes.Secondly,the reduction is performed through the neighborhood rough sets theory,and the importance of the condition attribute is compared to get the optimal reduction result.Finally,comparing the K-S test and the two feature selection methods based on the K-S test through experiments,this method can not only accurately select the tumor genes with significant ability of distinguishing,but also be efficient and feasible.

Keywords:K-S test;neighborhood rough sets;feature selection

[责任编校 陈留院]