ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 5,Sep. 2015,pp. 973－981
DOI:10.16337/j.1004-9037.2015.05.007

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于信噪比与邻域粗糙集的特征基因选择方法

徐久成[1,2]　李　涛[1,2]　孙　林[1,2]　李玉惠[1,2]

(1.河南师范大学计算机与信息工程学院,新乡,453007; 2.河南省高校计算智能与数据挖掘工程技术研究中心,新乡,453007)

摘　要:鉴于传统基因选择方法会选出大量冗余基因从而导致样本预测准确率较低,提出了一种基于信噪比与邻域粗糙集的特征基因选择方法(Signal noise ration and the neighborhood rough set, SNRS)。首先采用信噪比指标获得分类能力较强的预选特征子集;然后利用邻域粗糙集约简算法对预选特征子集进行寻优;最后采用不同的分类器对特征基因子集进行分类。通过实验表明,该方法能够克服传统分类算法精度不高的缺陷,并且能够在较少的特征基因下取得较高的分类精度,验证了该方法的可行性和有效性。

关键词:肿瘤基因表达谱;信噪比;邻域粗糙集;特征选择

中图分类号:TP18　　**文献标志码**:A

Feature Gene Selection Based on SNR and Neighborhood Rough Set

Xu Jiucheng[1,2], Li Tao[1,2], Sun Lin[1,2], Li Yuhui[1,2]

(1. College of Computer & Information Engineering, Henan Normal University, Xinxiang, 453007, China; 2. Engineering Technology Research Center for Computing Intelligence & Data Mining, Henan Province, Xinxiang, 453007, China)

Abstract: In view of that the traditional genetic selection method selects a large number of redundant genes, which leads to a lower sample forecast accuracy, a feature gene selection method is put forward based on the signal noise ration and the neighborhood rough set(SNRS). Firstly, the signal-to-noise ratio (SNR) index is used to obtain the primary feature subset which have a greater impact on classification. Secondly, the rough neighborhood intensive algorithm is used to optimize the primary feature subset. Finally, feature gene subset is classified by different classifier. Experiment results show that the proposed method can get a higher classification accuracy using less feature gene than the traditional ones, which verifies the feasibility and validity of the method.

Key words: gene expression profiles; signal-to-noise ratio; neighborhood rough set; feature selection

引　　言

随着大规模基因表达谱技术的发展,基因芯片为研究疾病的发病原理和临床疾病诊断提供了强有力的手段。肿瘤基因表达数据通常具有小样本、超高维的特点,且原始数据存在大量冗余基因和噪声,

基金项目:国家自然科学基金(61370169,61402153, 60873104)资助项目;河南省科技攻关重点(142102210056)资助项目;新乡市重点科技攻关计划(ZG13004)资助项目。
收稿日期:2015-05-29;**修订日期**:2015-07-02

因此在利用特征基因选择方法对新样本进行预测时,不仅花费大量时间,而且降低了分类精度[1-3]。因此如何识别对疾病有鉴别意义的特征基因或疾病相关基因是生物信息学的研究热点之一。

基因选择是从基因表达谱数据的所有属性中选择基因子集,且获得的基因具有较强的疾病识别能力[4-5]。基因排序法按照计分准则对每个基因计分,把分值较大的基因作为预选基因,基因分值越大,表明分类能力越强[6]。目前常用的特征基因计分准则主要包括信噪比指标(Signal-noise ratio, SNR),Fisher 判别(Fisher discriminant ratio, FDR)以及误分类阈值(Threshold number of misclassification score, TNM)等,其中信噪比指标应用最为广泛。基因排序法不依赖于具体的分类算法,并且得到的特征基因子集可有效避免"过拟合"现象,因此较适合用于基因表达数据[7-9]。目前常用的基因选择方法是过滤法和缠绕法[5]。其中,基于排序的过滤法如信噪比、信息增益等具有简单快速的特点,但是过滤法极有可能选择高相关的基因作为特征基因。这不仅会降低分类能力,而且也会增加额外的计算负担;而缠绕法将分类器预测的正确率评价基因子集,时间复杂度较高,且特征基因子集在其他分类器中的泛化能力较差。信噪比方法能有效地处理基因表达谱中的噪声问题,它根据基因对样本分类贡献大小的度量,过滤掉噪声基因,从而更有效识别出肿瘤基因。邻域粗糙集具有不需要对连续型数据离散化处理的特点,避免数据离散化所导致的信息损失。邻域粗糙集凭借其独特的优势,逐渐应用到生物信息学领域,并在肿瘤特征基因选取方面取得了一些较好的结果[10]。为了充分利用过滤法和缠绕法的优势,有效地去除无关基因和冗余基因,国内外专家提出了一些解决方法,文献[11]将遗传算法与支持向量机分类器相结合,把支持向量机的距离作为适应度函数评估特征基因的分类能力;文献[12]将随机森林用于基因选择和分类;文献[13]提出一种优化的邻域粗糙集的混合基因选择算法;文献[14]结合 K-means 和 Lasso 方法对基因表达谱数据进行特征选择和预测模型构建,取得较好的效果。虽然这些组合式特征选择方法在一定程度上提高了分类精度,但依然存在算法稳定性较差和特征子集规模较大的问题,如何在特征基因子集规模、分类能力和时间复杂度等多个目标下求得折中解是肿瘤基因分类领域的关键问题[15]。

鉴于肿瘤基因表达数据本身的特点,为了保证采用尽可能少的信息基因获得尽可能高的样本分类率,同时降低算法的时间复杂度,本文提出一种基于信噪比与邻域粗糙集的特征基因选择方法。首先选取信噪比值较大的基因作为预选特征基因子集;然后利用邻域粗糙集约简算法对预选特征子集进行寻优;最后通过仿真实验验证该方法的有效性和可靠性。

1 基本概念

1.1 信噪比

信噪比是一种简单高效的排序法[4]。在进行基因选择过程中,首先采用的信噪比指标在原始特征空间过滤无关基因,得到与类别属性相关性较高的基因,信噪比为

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

式中:$\mu_+(g_i)$和$\mu_-(g_i)$分别表示第 i 个基因 g_i 在正类和负类的平均表达值;而 $\delta_+(g_i)$ 和 $\delta_-(g_i)$ 分别表示第 i 个基因 g_i 在两类中的标准差。式(1)用来衡量每个基因的重要性,信噪比值越大,说明该基因的重要性越强。

1.2 邻域决策系统

在经典粗糙集基础上,文献[16]提出了邻域粗糙集模型,该模型能够直接处理连续型数据,不需要对连续型数据进行离散化处理,从而避免了离散化过程中的信息损失。下面给出邻域粗糙集模型的相关概念和性质[12,17,18]。

定义 1 设 $U=\{u_1,u_2,\cdots,u_n\}$ 为样本集;C 为条件属性集;D 为决策属性集;N 为由 C 产生的一簇邻域关系,称 $\langle U,C\cup D,N\rangle$ 为邻域决策系统。

定义 2 在邻域决策系统 $\langle U,C\cup D,N\rangle$ 中,D 将 U 划分成 N 个等价类:X_1, X_2, X_3,$\cdots$,X_N,$\forall B\subseteq C$ 生成 U 上的邻域关系 N_B,则决策属性 D 关于 B 的邻域下近似、上近似分别定义为

$$\underline{N_B}D=\{\underline{N_B}X_1,\underline{N_B}X_2,\cdots,\underline{N_B}X_N\} \tag{2}$$

$$\overline{N_B}D=\{\overline{N_B}X_1,\overline{N_B}X_2,\cdots,\overline{N_B}X_N\} \tag{3}$$

式中:$\underline{N_B}D=\{x_i\mid\delta_B(x_i)\subseteq X,x_i\in U\}$;$\overline{N_B}D=\{x_i\mid\delta_B(x_i)\cap X\neq\varnothing,x_i\in U\}$,决策边界为 $BN(D)=\overline{N_B}D-\underline{N_B}D$。

定义 3 在邻域决策系统 $\langle U,C\cup D,N\rangle$ 中,称 $\gamma_B(D)=\mathrm{Card}(\underline{N_B}D)/\mathrm{Card}(U)$ 为决策属性 D 对条件属性 $B\subseteq C$ 的依赖度。

定义 4 在邻域决策系统 $\langle U,C\cup D,N\rangle$ 中,$\forall a\in B\subseteq C$,若 $\gamma_B(D)>\gamma_{B-a}(D)$,称 a 在 B 中相对决策属性 D 是必要的,否则是不必要的。

定义 5 在邻域决策系统 $\langle U,C(D,N\rangle$ 中,若 $B\subseteq C$ 满足:(1) $\gamma_B(D)=\gamma_C(D)$;(2) $\forall a\in B$,$\gamma_{B-a}(D)<\gamma_B(D)$,则称 B 是 C 的一个相对约简。

定义 6 在邻域决策系统 $\langle U,C\cup D,N\rangle$ 中,若 $B\subseteq C$,$a\in C-B$,则 a 关于属性子集 B 的重要度定义为 $SIG(a,D,B)=\gamma_{B\cup a}(D)-\gamma_B(D)$。

2 特征基因选择方法

2.1 过滤无关基因

信噪比方法简单高效且能有效处理基因表达谱中的噪声问题,而 Relief 算法具有计算复杂度小和考虑属性间相关性的特点。本文利用信噪比去除基因表达数据中的无关基因,按照信噪比值的大小对全部基因进行降序排列,将排好的基因变量以 0.2 为单位划分到不同的区间,分别为(0,0.2],(0.2,0.4],(0.4,0.6],(0.6,0.8],(0.8,1]。因此,原始基因表达谱数据集被划分为 5 个特征基因子集,每个区间的基因均可作为预选特征基因子集。在此基础上,采用经典的 Relief 算法给出基因的分类权重,过滤掉各区间权重较小的基因。

为了选取噪声较少且与分类高相关的预选特征基因子集,本文只选取区间值最大的基因子集作为较优的候选基因子集。由于通过式(1)计算本文的 3 个数据集信噪比值在区间(0.8,1]的基因数目为零,为了保持算法的整体性能,不考虑 Prostate 数据集信噪比大于 0.8 的基因。若基因的信噪比越大,表明该基因对分类的影响越大,因此选取各数据集在(0.6,0.8]区间内的基因作为候选特征基因子集。但是,候选特征基因子集中往往存在冗余基因,这不仅增加额外计算负荷,而且导致错误的分类结果。因此本文采用邻域粗糙集进一步剔除冗余基因,从候选特征基因集合中获取较优的特征基因子集。

2.2 特征基因选择方法

邻域粗糙集能够直接处理连续型数据,它可以直接应用于特征基因的提取,避免了一定程度上的信息丢失,使得所选取的特征基因子集能最大限度地保持原数据集的分类能力。下面给出有关算法定义[16-17]。

定义 7 在给定的 N 维实数空间 Ω 中,$\mathbf{R}$ 为实数集,$\mathbf{R}^N$ 为 N 维实数向量空间,$\boldsymbol{\Delta}:\mathbf{R}^N\times\mathbf{R}^N\to\mathbf{R}$,则称 $\boldsymbol{\Delta}$ 为 $\mathbf{R}^N$ 上的一个度量,若 $\boldsymbol{\Delta}$ 满足以下条件:(1) 对 $\forall\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathbf{R}^N$,有 $\boldsymbol{\Delta}(\boldsymbol{x}_1,\boldsymbol{x}_2)\geqslant 0$,其中当且仅当 $\boldsymbol{x}_1=\boldsymbol{x}_2$ 时等号成立;(2) 对 $\forall\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathbf{R}^N$,有 $\boldsymbol{\Delta}(\boldsymbol{x}_1,\boldsymbol{x}_2)=\boldsymbol{\Delta}(\boldsymbol{x}_2,\boldsymbol{x}_1)$;(3) 对 $\forall\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\in\mathbf{R}^N$,有 $\boldsymbol{\Delta}(\boldsymbol{x}_1,\boldsymbol{x}_3)\leqslant\boldsymbol{\Delta}(\boldsymbol{x}_1,\boldsymbol{x}_2)+\boldsymbol{\Delta}(\boldsymbol{x}_2,\boldsymbol{x}_3)$,则称 $(\Omega,\boldsymbol{\Delta})$ 为度量空间,其中 $\boldsymbol{\Delta}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 为距离函数,表示元素 $\boldsymbol{x}_i$ 和元素 $\boldsymbol{x}_j$ 之

间的距离。

距离计算函数有曼哈顿距离函数、欧几里德距离函数、P 范式距离函数，本文算法从特征选择方法模型泛化的角度考虑，采用欧几里德距离函数，它能够在一定程度上防止过拟合，提升模型的泛化能力。

假设一个含有 K 个样本 T 个属性的基因数据集，直接利用邻域粗糙集的向前属性约简算法剔除冗余基因时，算法的计算代价较大，时间复杂度为 $O(T^2 \times K\log K)$[19]。当属性的邻域半径一定时，随着属性集 B 中属性个数的增加，会增加样本的误判率，从而降低分类正确率。本文提出基于信噪比与邻域粗糙集的特征基因选择算法，该算法可以有效去除大量的无关基因，减少邻域粗糙集约简过程的时空消耗，同时也减少分类器的训练时间，具体算法如下。

输入：基因数据集 Set$=(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, y)$，邻域决策系统 NDS$=\langle U, C \cup D, N\rangle$，计算属性邻域半径的参数 r 及属性的重要度下限参数 λ。

输出：特征基因集合 S。

(1) 对 Set 每个属性列进行标准化处理；

(2) 根据式(1)计算每个基因变量的信噪比值；

(3) 根据信噪比的大小对 G_{list} 进行升序排序；//G_{list} 表示通过信噪比排序后的基因列表

(4) 将信噪比值在区间[0.6, 0.8]的标准化基因数据生成新的矩阵 $\boldsymbol{A}_{l\times t}$；//$l$ 为样本数，t 为属性数

(5) 利用 Relief 算法过滤掉各区间权重较小的基因；

(6) 将矩阵 $\boldsymbol{A}_{l\times t}$ 中所有属性列组成特征基因集合 S_A；//S_A 为已过滤掉无关基因的集合；

(7) 初始化约简集合 $red=\varnothing$；

(8) 对 S_A 中的 $a_i \in S_A - red$；//a_i 表示特征基因集合 S_A 的属性列，$i=1, 2, \cdots, t$；

计算 a_i 的正域 $pos^U_{red+a_i}(D)$ 及其重要度 SIG，若某基因的重要度 SIG 为零，说明该基因为冗余基因

(9) 获取属性 a_i 的最大的正域 $pos_k(D)$；//通过最大的正域 $pos_k(D)$ 计算属性的重要度；

(10) 判断重要度 SIG 是否大于设定的下限 λ；

(11) 若 $SIG \leqslant \lambda$，记录 k 值，$red = red + a_k$，$S = S - pos_k$，返回(8)；//通过 k 记录对应的属性列号

(12) 若 $SIG > \lambda$，输出约简结果 red；

(13) 根据 red 对应的属性列，获取较优的特征基因集合 S；

(14) 结束。

假设一个含有 K 个样本 T 个属性的基因数据集，经过信噪比去除无关基因后获得 M 个特征基因，平均选择一个特征基因要向正域集合中添加 K/M 个样本，则计算该数据集邻域时间复杂度为 $O(K\log K)$。由于第 1 个特征基因确定正域的时间复杂度为 $T\times K\log K$，第 2 个特征基因的时间复杂度为 $(T-1)\times(K-K/M)\log(K-K/M)$，则第 M 个特征基因的时间复杂度为 $(T-M+1)\times(K/M)\log(K/M)$，经计算得到 SNRS 算法的时间复杂度为 $M\times T\times K\log K$，因为 $M \ll T$，所以本文算法的时间复杂度小于 $O(T^2\times K\log K)$。由以上分析可知，该算法通过约简过滤掉信噪比值小的基因，从而减小了时间复杂度。

3 实验分析

3.1 实验数据和实验环境

为了验证算法的有效性，本文在 Leukemia，Colon，Lung 和 Prostate 4 个公开的基因表达谱数据集上仿真实验。数据集从 http://datam.i2r.a-star.edu.sg/datasets/krbd/下载，具体数据集描述见表 1。实验中所用的计算机配置为酷睿 i5-3470，3.20 GHz，2 GB 内存，所有仿真都在 Matla-

bR2010a 和 Weka3.6.11 中实现,并构建朴素贝叶斯、Libsvm 和决策树 C4.5 三种分类模型,其中 Libsvm 的核函数设置为线性核函数,C4.5 用于修剪的置信因子设置为 0.25。所有实验都采用 k 折交叉验证方法,其中 k 均取值为 10。

表 1 实验数据集描述

Table 1 Describe experimental dataset

序号	数据集名称	基因数量	样本数量(正类/负类)	类别数
1	Leukemia	7 129	72(25/47)	2
2	Colon	2 000	62(40/22)	2
3	Lung	2 880	39(15/24)	2
4	Prostate	12 600	102(52/50)	2

3.2 实验结果分析

根据信噪比值的大小,将基因变量分布在 4 个区间,为了直观表示,图 1～4 分别给出了 4 个数据集信噪比值相应的区间分布。由图 1～4 可知,本文实验将全部基因信噪比值分为 4 个区间:(0,0.2],(0.2,0.4],(0.4,0.6],(0.6,0.8],4 个数据集在相应区间内的基因个数分别为{4 973, 1 796, 334, 26},{1 481,509,10,0},{1 727,969,174,10},{6 976,5 156,411,49}。可知大部分基因的信噪比值都较小,如 Lung 数据集的基因数目为 2 880,其中有 1 727 个基因的信噪比值小于或等于 0.2;Prostate 数据

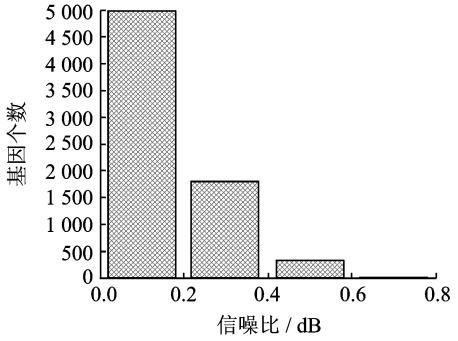


图 1 Leukemia 数据集的区间分布图

Fig. 1 Leukemia dataset interval distribution

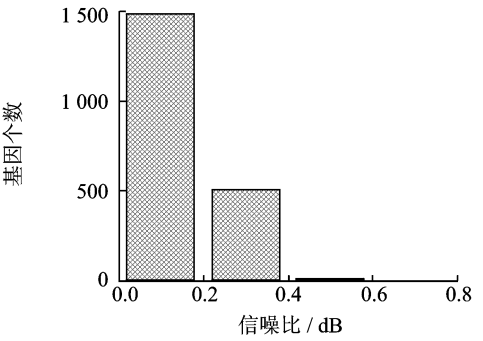


图 2 Colon 数据集的区间分布图

Fig. 2 Colon dataset interval distribution

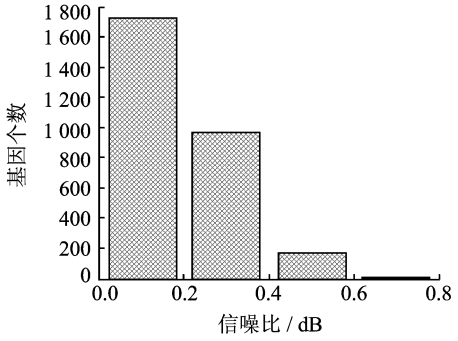


图 3 Lung 数据集的区间分布图

Fig. 3 Lung dataset interval distribution

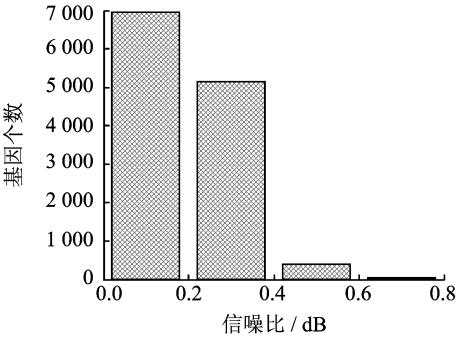


图 4 Prostate 数据集的区间分布图

Fig. 4 Prostate dataset interval distribution

集的基因数目为 12 600,其中有 6 976 个基因的信噪比值小于或等于 0.2。表明这些基因难以区分类别,可做无关基因处理,只有为数不多的基因与样本的分类有密切相关。为了有效获取特征基因子集,本文只将信噪比值分布在区间(0.6, 0.8]内基因变量作为预选的特征基因子集。由于 Colon 数据集在区间(0.6, 0.8]的基因变量为零,所以将它在区间(0.4, 0.6]内的基因变量作为预选的特征基因子集。因此,以上 4 个数据集通过信噪比去除无关基因和 Relief 算法去除权重较小的基因,最终获得预选的特征基因子集数目分别为 21,8,9 和 46。

利用邻域粗糙集剔除预选特征基因子集中的冗余基因,本文对计算邻域半径参数 r 和重要度下限进行了优化,经过多次试验比较,r 的取值[3.5, 4.5]较为合适,实验中将 r 取值为 4,重要度下限取值越小越好,因此取值为 0.001。学习分类算法中的朴素贝叶斯具有训练速度较快的特点,支持向量机能避免"维数灾难",具有较好的鲁棒性,而 C4.5 具有处理不完整数据及分类规则易理解的特点。为了证实本文算法在分类模型优于其他的特征基因选择方法,实验采用朴素贝叶斯、Libsvm、决策树 C4.5 三种学习算法验证各自的分类性能,如图 5~7 所示。

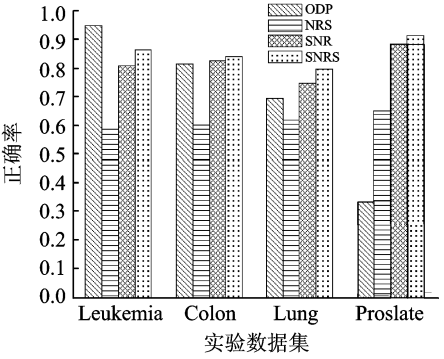


图 5 Naive Bayesian 在数据集上的分类性能

Fig. 5 Classification performance of Naive Bayesian on data sets

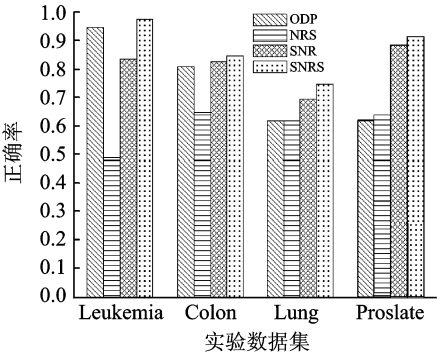


图 6 Libsvm 在数据集上的分类性能

Fig. 6 Classification performance of Libsvm on data sets

在图 5~7 中,ODP(Original data processing)表示为直接对原数据集分类的方法;NRS(Neighborhood rough set)表示为仅采用邻域粗糙集方法;SNR 表示为只采用信噪比方法;SNRS 表示为本文算法即采用基于信噪比与邻域粗糙集的方法。由图 5~7 可知,针对某一特定的数据集,不同的特征基因选择方法在 3 种分类器上表现出不同的分类性能。通过本文算法与其他方法相比较,基于信噪比与邻域粗糙集的算法的分类正确率相对较高。例如在 Prostate 数据集中,本文方法在朴素贝叶斯分类器、Libsvm 分类器、C4.5 分类器得到分类正确率分别为 91.176 5%,91.176 5%,90.196 1%,明显高出其他方法的分类正确率。但是在 Leukemia 数据集中,本文算法在 Naive Bayesian 分类器的分类正确率为 86.111 1%低于 ODP 方法在该分类器上 94.736 8%的分类正确率,这表明在利用本文方法去除无关基因和冗余基因时,错误地去除了对分类影响较大的基因变量,最终影响了样本分类的正确率。但是该算法在其余 3 个数据集上都表现出良好的性能。因此,本文算法在整体上能够获取高度相关、低度冗余的特征基因子集,并且有效提高了

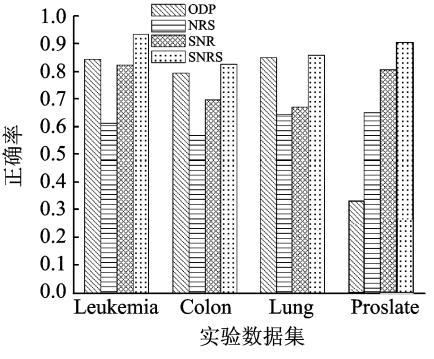


图 7 C4.5 在数据集上的分类性能

Fig. 7 Classification performance of C4.5 on dataset

特征基因分类算法的正确率。

由表 2 可知,ODP 算法虽然可获得较高的分类精度,但是特征基因规模过于庞大;NRS 算法可有效地去除无关基因,但是在去除冗余基因过程中也剔除了与分类相关的特征基因,从而导致分类精度较低;SNR 算法可获取较少的特征基因子集,并且分类性能也相对较好。而一个较为理想的特征基因选择方法不仅能获得较少的特征基因子集,同时也具有较高的分类精度。基于信噪比与邻域粗糙集的算法的分类精度相对其他算法最高,特征基因个数也相对较少。例如在 Leukemia 数据集上,获得 4 个特征基因相对其他方法最少,与此同时,分类性能也达到 97.36%的正确率。从表 2 可知,虽然算法 SNRS 约简后的特征基因个数不少于算法 NRS 约简前的特征基因个数,但两者仅差 1～3 个特征基因,同时 SNRS 方法在 4 个数据集上都获得最高分类精度。例如在 Prostate 数据集上,SNRS 算法获得 5 个特征基因数目,虽然比 NRS 算法中获得 4 个特征基因多 1 个特征基因,但是分类精度已高达 91.18%。由表 3 可知,LASSO 方法可获得较优的分类精度,但其时间复杂度高达 $O(PT^3)$;NRS 方法可有效减少特征基因个数但其分类精度最低;MRMR 方法的分类精度略高,但其特征基因个数较多,时间复杂度也较高。与其他 3 个经典特征基因选择方法相比,本文方法在特征基因子集规模与分类精度上均取得较好的结果,且时间复杂度也较低,综合性能较强。例如在 Leukemia 数据集上,本文方法获得 4 个特征基因均不多于其余 3 种方法,分类精度高达 97.36%,略低于 LASSO 方法的 98.61%。

表 2 各种算法在不同数据集上的特征基因个数和最优分类性能的实验对比

Table 2 Experimental contrast of all kinds of algorithms on different data sets feature gene number and optimal classification performance

数据集	ODP		NRS		SNR		SNRS	
	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/%
Leukemia	7 129	94.44	4	61.11	26	83.33	4	97.36
Colon	20 008	81.10	5	64.52	10	82.26	6	82.26
Lung	2 880	84.62	3	64.10	10	74.36	6	85.44
Prostate	12 600	61.90	4	64.71	49	88.24	5	91.18

表 3 不同特征基因选择方法优分类性能和时间复杂度的实验对比

Table 3 Experimental comparison of classification and time complexity of different feature gene selection methods

数据集	LASSO		NRS		MRMR		本文方法	
	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/%
Leukemia	23	98.61	4	61.11	28	89.06	4	97.36
Colon	5	88.71	5	64.52	54	79.86	6	82.26
Lung	8	99.45	3	64.10	36	84.61	6	85.44
Prostate	63	96.08	4	64.71	79	92.15	5	91.18
时间复杂度	$O(PT^3)$		$O(T^2K\log K)$		$O(T^2)$		$O(MTK\log K)$	

由实验结果可知,基于信噪比与邻域粗糙集的算法能够选择出较少的特征基因,通过该方法获取的基因数目均不高于 6 个特征基因,最少的只达到 4 个特征基因。在特征基因子集规模如此小的情况下,本文方法在整体性能上均高于其他 3 种基因选择方法,从而证明基于信噪比与邻域粗糙集的算法能选择出高信息含量的基因,同时也能减少了选择基因子集的冗余性。总之,本文算法能选出基因数量较少且分类能力较强的特征基因子集,解决了基因表达谱数据高维数、高冗余问题,提高了分类模型的精度和泛化能力。

4 结束语

从 DNA 微阵列中选择分类能力且数量较少的特征基因子集是生物信息学领域研究的一个重要问题。本文针对目前特征基因选择算法分类精度较差和时间复杂度较高的问题，提出了一种基于信噪比与邻域粗糙集的特征基因选择方法。该方法分为两个过程，利用信噪比指标衡量基因的重要性，并划分不同区间，以过滤无关基因；采用邻域粗糙集进行冗余基因的剔除。实验结果表明，该方法能够选择出具有高分辨率且特征基因数目较少的基因子集，并且提高了算法的分类精度并且降低了时间复杂度。本文提出的算法研究了单个特征基因类间区分度，在特征基因选择过程中考虑多个特征基因对分类的联合贡献及如何提高算法时间效率将是下一步的研究工作。

参考文献：

[1] 李霞，张田文，郭政. 一种基于递归分类树的集成特征基因选择方法[J]. 计算机学报，2004，27(5)：675-682.
Li Xia, Zhang Tianwen, Guo Zheng. An integrated feature gene selection based on the recursive classification tree method [J]. Chinese Journal of Computers, 2004, 27(5): 675-682.

[2] 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.
Xu Feifei, Miao Duoqian, Wei Lai. Tumor classification feature gene selection based on fuzzy rough sets[J]. Computer Science, 2009, 36(3): 196-200.

[3] 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.
Wang Jingqi, Xu Linli. Semi-supervised feature selection and clustering for multi-view data[J]. Journal of Data Acquisition and Processing, 2015, 30(1): 106-116.

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

[5] 周昉，何洁月. 生物信息学中的基因芯片的特征选择技术综述[J]. 计算机科学，2007，34(12)：143-150.
Zhou Fang, He Jieyue. Survey of the gene selection technologies based on microarray in bioinformatics[J]. Computer Science, 2007, 34(12): 143-150.

[6] 黄德双. 基因表达谱数据挖掘方法研究[M]. 北京：科学出版社，2009.
Huang Deshuang. Gene expression profile data mining methods[M]. Beijing: Science Press, 2009.

[7] Zhao Y H, Yu X J, Wang G R, et al. Maximal subspace coregulated gene clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20 (1): 83-98.

[8] 刘金勇，郑恩辉，陆慧娟. 基于聚类与微粒子群优化的基因选择方法[J]. 数据采集与处理，2014，29(1)：84-89.
Liu Jinyong, Zheng Enhui, Lu Huijuan. Gene selection based on clustering method and particle swarm optimazition[J]. Journal of Data Acquisition and Processing, 2014, 29(1): 84-89.

[9] 李建更，郭庆雷，贺益恒. 时序基因表达缺失值的加权上相回归估计算法[J]. 数据采集与处理，2013，28(2)：137-140.
Lin Jiangeng, Guo Qinglei, He Yiheng. Double weighted regression estimation for missing values in time series gene expression data[J]. Journal of Data Acquisition and Processing, 2013, 28(2): 137-140.

[10] 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，2014，35(11)：2529-2532.
Xu Jiucheng, Xu Tianhe, Sun Lin, et al. Feature selection for cancer classification based on neighborhood rough set and particle swarm optimization[J]. Journal of Chinese Computer Systems, 2014, 35(11): 2529-2532.

[11] Chen X W. Margin-based wrapper methods for gene identification using microarray[J]. Neurocomputing, 2006, 69(18): 2236-2243.

[12] Ramón D U, Sara A A. Gene selection and classification of microarray data using random forest[J]. BMC Bioinformatics, 2006(7): 3-4.

[13] 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.
Cheng Tao, Hong Zenglin, Deng Fangan. Hybrid gene selection algorithm based on optimized neighborhood rough set[J]. Computer Science, 2014, 41(10): 291-294.

[14] Ma Shuangge, Song Xiao, Huang Jian. Supervised group Lasso with applications to microarray data analysis[J]. BMC Bioinformatics, 2007(8): 60.

[15] 王楠，欧阳丹彤. 基于模型诊断的抽象分层过程[J]. 计算机科学，2011，34(2)：384-394.

Wang Nan, Ouyang Dantong. Hierarchical abstraction process in model-based diagnosis[J]. Computer Science, 2011, 34(2): 384-394.

[16] 胡清华,于达仁. 基于邻域粒化和粗糙逼近的属性约简[J]. 软件学报, 2008, 15(3):121-125.
Hu Qinghua, Yu Daren. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of Software, 2008, 15(3): 121-125.

[17] 张文修,仇国芳. 粗糙集属性约简的一般理论[J]. 中国科学:技术科学, 2005,35(12):1304-1313.
Zhang Wenxiu, Qiu Guofang. The general theory of rough set attribute reduction[J]. Scientia Sinica Technologica, 2005,35(12):1304-1313.

[18] Chen T. Classification algorithm on gene expression profile of tumor using neighborhood rough set and support vector machine[J]. Advanced Materials Research, 2014, 850:1238-1242.

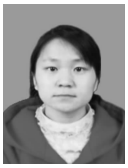
[19] 谢娟英,李楠,乔子芮. 基于邻域粗糙集的不完整决系统特征选择算法[J]. 南京大学学报:自然科学, 2011, 47(4): 384-390.
Xie Juanying, Li Nan, Qiao Zirui. Feature subset selection algorithms for incomplete decision systems based on neighborhood tough sets[J]. Journal of Nanjing University:Natural Sciences, 2011, 47(4): 384-390.

作者简介:

徐久成(1964-),男,教授,博士生导师,研究方向:粒计算、粗糙集、数据挖掘和生物信息学等, E-mail: xjch3701@sina.com。

李涛(1990-),男,硕士研究生,研究方向:数据挖掘、粗糙集和生物信息学。

孙林(1979-),男,讲师,研究方向:粒计算、粗糙集和数据挖掘。



李玉惠(1988-),女,硕士研究生,研究方向:粒计算和图像检索。