基于优化的邻域粗糙集的混合基因选择算法

陈 涛[1,2] 洪增林[1] 邓方安[2]
(西北工业大学自动化学院 西安 710072)[1] (陕西理工学院数学与计算机科学学院 汉中 723000)[2]

摘 要 DNA微阵列技术可以同时检测细胞内成千上万的基因的活性,被广泛应用于重大基因疾病的临床诊断。然而微阵列数据通常具有高维小样本特点,且存在大量噪声和冗余基因。为了进一步提高微阵列数据分类性能,提出一种特征基因混合选择算法。首先采用ReliefF算法剔除大量无关基因,获得特征基因候选子集;然后采用基于差分进化算法优化的邻域粗糙集模型实现特征基因选择;最后利用支持向量机进行分类,以验证算法的有效性。仿真实验结果表明,该算法能用尽可能少的特征基因来获得更高的分类精度,既增强了算法的泛化性能,又提高了时间效率,而且对致病基因的临床诊断有着重要的参考意义。

关键词 特征基因选择,ReliefF算法,邻域粗糙集模型,差分进化算法
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.10.061

Hybrid Gene Selection Algorithm Based on Optimized Neighborhood Rough Set

CHEN Tao[1,2] HONG Zeng-lin[1] DENG Fang-an[2]
(School of Automation, Northwestern Polyechnical University, Xi'an 710072, China)[1]
(School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong 723000, China)[2]

Abstract DNA microarray technique can detect tens of thousands of gene activity in cells, which has been widely used in clinical diagnosis. However, microarray data has high dimension, small sample, a lot of noise and redundant genes. In order to further improve the classification performance, this paper proposed a hybrid gene selection algorithm. Firstly, using ReliefF algorithm to eliminate a lot of irrelevant genes, the feature genes candidate set was obtained. Then the optimized neighborhood rough set model based on differential evolution algorithm was used to select feature genes. At last the validity of the algorithm was verified using support vector machine as classifier. The simulation results show that the algorithm can obtain higher classification accuracy with less feature gene, and it not only enhances the generalization performance of the algorithm, but also improves the time efficiency.

Keywords Feature gene selection, ReliefF algorithm, Neighborhood rough set model, Differential evolution algorithm

1 引言

DNA微阵列技术是上世纪末分子生物学领域的一项重大技术,它能够在一次实验中同时检测细胞中成千上万的基因的表达水平,从分子层面对疾病尤其是恶性肿瘤诊断、致病机理研究提供了强有力的支持[1,2]。然而微阵列数据往往仅含几十或上百个样本,而基因(即特征或属性)数却达到几千甚至上万,这就导致了样本数与基因数之间不平衡,造成"维数灾难"问题。而且微阵列数据中包含许多不相关的以及冗余的基因,这些基因的存在,不仅增加特征空间的维数、降低学习的效率,而且还增加了噪声数据干扰的可能,从而干扰算法的学习过程,最终影响分类模型的构造和结果。因此,为了降低这种不利因素的影响,减少特征空间的维数,不相关及冗余基因应从数据中剔除,以减少噪声数据的干扰,有效地提高学习算法的效率和性能,并避免样本数较少时出现的"过拟合"现象,基因选择是解决以上问题的一种有效方法[3,4]。

Filter法是常用的基因选择方法,它根据基因数据的内在特性,构造独立于分类器的评价指标来度量基因与分类任务的关联程度。Golub等提出信噪比(SNR)准则[5];Arfin提出T-statistic[6];Usher等改进T-statistic,提出SAM法[7];Efron等提出wilcoxon统计量[8]等。Filter法原理简单、计算方便,具有线性时间复杂度,运算速度快。但大部分Filter法要求数据满足正态分布,且没有考虑基因之间的相互作用,无法剔除冗余基因,不能发现基因相互影响引起的分类能力的变化,且因为没有分类器参与,所以选择的基因往往得不到较好的分类效果。

在众多Filter法中,Relief[9]及ReliefF[10]是公认的效果较好的方法,其核心思想是以属性区分"相近"样本的能力来作为其评估属性重要性的指标,且由此给出属性的分类权重。该算法对数据的分布没有特殊要求,可以处理连续型和离散型数据,而且计算的复杂度较小,并且在某种程度上考虑了属性之间的相关性,是一种约简无关属性的有效方法,相比其他

收稿日期:2013-12-02 返修日期:2014-03-11 本文受国家自然科学基金(81160183,11305097),陕西省教育厅科研计划项目(147JK1132),陕西理工学院基金(SLGKY13-41,SLGKY13-44)资助。

陈 涛(1979—),男,博士生,副教授,CCF会员,主要研究方向为数据挖掘、生物信息学,E-mail:ct79hz@126.com;洪增林(1963—),男,博士,教授,主要研究方向为系统工程;邓方安(1964—),男,博士,教授,主要研究方向为智能信息处理。

Filter 法具有更好的分类性能。

粗糙集[11,12]是波兰学者 Pawlak 提出的一种刻画由不精确信息描述的分类问题中的不一致性的数学工具，它的一个重要应用是属性约简，即找到一个最小的、与全部属性具有相同区分能力的属性子集，通过粗糙集的属性约简可以改善运算速度、存储以及精度等。但它在约简之前，需要对连续数据进行离散化，而离散化过程会引入量化误差，且改变了数据的本质结构，所以在一定程度上影响原属性集的信息表达，由于信息的损失及其没有得到充分利用，导致分类效果降低。另外，离散化技术中的断点选择也是一个很困难的问题。邻域粗糙集模型[13,14]是对经典粗糙集的改进，可以直接处理连续型数据，而不需要离散化，这样就避免了离散化所带来的信息损失，可以更有效地提取原属性集的信息，消除决策属性表中的无关和冗余属性。

本文首先利用 ReliefF 算法对微阵列数据进行基因初选以剔除不相关的基因，获得基因候选子集；由于邻域粗糙集的半径是影响该模型性能的重要因素，因此利用差分进化算法[15]的全局寻优能力对半径进行优化，获得最优半径；最后采用优化的邻域粗糙集进行基因选择，算法中采用支持向量机作为分类器进行分类。通过在 3 个标准微阵列肿瘤数据集进行仿真实验来验证本文算法的有效性。

2 基于 ReliefF 算法的特征基因初选

Relief 算法[9]由 Kira 于 1992 年提出，该算法借用邻近学习算法的思想，从训练样本中随机选取 m 个样本，通过所选取的样本与属于同类和不同类的两个最近邻样本的差异，求出每个样本各特征与类的相关性，然后取平均值作为每个特征的权值，获得每个特征与类的相关性度量。但是 Relief 只能用于两分类问题。1994 年，Kononenko 对 Relief 算法进行扩展得到 ReliefF 算法[10]，其可以直接处理多分类问题，且增加了噪声信息的容忍机制。它从同类和不同类中各取 k 个最近邻样本，求其平均值得到每个特征的权值，从而获得每个样本中各特征与类的相关性。ReliefF 算法采用 k 个同类最近样本和每个异类中 k 个最近样本计算的类概率加权距离来衡量样本的可分性，不受个别噪声样本的影响，具有很强的稳健性。

基于 ReliefF 的属性约简算法描述如下：

输入：训练样本集 $X=(x_1,x_2,\cdots,x_n)$

输出：特征权值向量 $W=(w_1,w_2,\cdots,w_m)$

1. 初始化特征权值向量：$W=(0,0,\cdots,0)$；

2. For i=1 to n

(1)在训练集中随机挑选一个样本 x_i；

(2)在 x_i 同类中选出与 x_i 最近的 k 个样本 H_j；

(3)For eachclass C≠class(x_i)

(a)在与 x_i 的每个不同类中，各选出与 x_i 最近邻的 k 个样本 M_j；

(b) For g=1 to m

$$W(g)=W(g)-\left(\sum_{j=1}^{k}\mathrm{diff}(g,x_i,H_j)\right)/kn+\sum_{C\neq \mathrm{class}(x_i)}\left(\frac{P(C)}{1-P(\mathrm{class}(x_i))}\left(\sum_{j=1}^{k}\mathrm{diff}(g,x_i,M_j)\right)/kn\right)$$

End；

End；

End；

3) 输出 W。

其中，$P(C)=num(C)/n$，$num(c)$ 是 C 类样本；$diff(g,x,y)=|value(g,x)-value(g,y)|/\max(g)-\min(g)$，$value(g,x)$ 是样本 x 在基因 g 的值，$\max(g)$、$\min(g)$ 分别为属性 g 在所有样本中的最大值与最小值。

3 基于优化的邻域粗糙集的特征基因选择

3.1 邻域粗糙集

粗糙集[11]用于解决不完整、不精确信息的数学工具，它不需要任何先验信息，仅使用数据本身的内在信息就能从中发现隐含知识，揭示潜在规律，从而对不完整不精确数据进行有效处理。但粗糙集方法首先需要将连续型数据离散化，而离散化处理会丢失对象之间的距离信息、大小序信息、邻域结构以及模糊隶属信息等，从而导致学习到的模型不能很好地反映原始数据的真实结构。在经典粗糙集基础上，胡清华等提出邻域粗糙集模型[13,14,16-19]，其能够直接处理连续型数据，不再需要对连续型数据进行离散化处理，避免离散化过程中的信息损失，直接用于知识约简。下面给出了邻域粗糙集模型的相关概念和性质[13,14]。

定义 1 设 $U=\{u_1u_2,\cdots,u_n\}$ 是样本集，C 是条件属性集，D 是决策属性集，N 是由 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：$X_1,X_2,\cdots,X_N$，$\forall B\subseteq C$ 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：

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

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

其中，$\underline{N_B}X=\{x_i\mid\delta_B(x_i)\subseteq X,x_i\in U\}$，$\overline{N_B}X=\{x_i\mid\delta_B(x_i)\cap X\neq\varnothing,x_i\in U\}$。

决策边界定义为：$BN(D)=\overline{N_B}D-\underline{N_B}D$。

定义 3 在邻域决策系统 $\langle U,C\cup D,N\rangle$ 中，称 $\gamma_B(D)=Card(N_BD)/Card(U)$ 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 $\langle U,C\cup D,N\rangle$ 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，$B\subseteq C,a\in C-B$，则 a 关于属性子集 B 的重要度定义为：$SIG(a,D,B)=\gamma_{B\cup a}(D)-\gamma_B(D)$。

基于邻域粗糙集的向前属性约简算法描述如下(FARNRS)：

输入：邻域决策系统 $\langle S,C\cup D,N\rangle$ 及邻域半径 δ；

输出：约简的 red。

1. red=∅；

2. For each $a_i\in C-$red

计算 $\gamma_{red\cup a_i}(D)=|POS_{B\cup a_i}(D)|/|U|$，及 $SIG(a_i,D,red)=\gamma_{red\cup a_i}(D)-\gamma_{red}(D)$；

End

3. 选择 a_k，使其满足；

$SIG(a_k,D,red)=\max(SIG(a_i,D,red))$；

4. if $SIG(a_k, D, red) > 0$

$red = red \cup a_k$;

Go to 2

Else

5. 输出 red.

3.2 基于差分进化算法的邻域粗糙集半径优化

差分进化算法(Differential Evolution, DE)是 Stom R 等在 1995 年提出的一种基于群体差异的启发式随机并行搜索算法。与传统进化算法相比,DE 采用实数编码,降低遗传操作的复杂性,特别是其特有的记忆功能可以动态跟踪当前搜索情况,具有原理简单、受控参数少、全局收敛能力较好及鲁棒性等优点。大量研究表明[15,20],差分进化算法比遗传算法、粒子群算法、蚁群算法等具有更快的收敛速度,更容易获得全局最优解。

邻域粗糙集模型是由实数空间中的每一个点形成一个 δ 邻域,而 δ 邻域则成为描述空间中任意概念的基本信息粒子。邻域粗糙集中分类边界的大小不仅与分类问题的特征空间有关,而且与分析的粒度,即邻域的大小有关。邻域的大小反映了进行分类时是在多大的粒度下来区分对象,决定了在分类边界附近区域参加训练样本的数量,所以邻域半径 δ 是影响邻域粗糙集模型性能的重要因素。

目前,邻域半径的取值没有一个统一的标准,且其大小往往与研究对象有关,常常是通过实验进行观察,这样效率较低,且往往无法得到最优。本文基于差分进化算法的全局寻优能力来优化邻域半径。

基于差分进化算法的邻域粗糙集半径优化算法描述如下:

输入:训练集 S、测试集 T、种群规模 NP、变异算子 F、交叉算子 CR、最大迭代次数 T;

输出:最优半径 δ。

1. 种群初始化

利用 $x_{i,0} = x_{low} + rand[0,1] \cdot (x_{upper} - x_{low})(i = 1,2,\cdots,NP)$ 随机产生 NP 个长度为 1 的实数向量,每一个向量代表种群中的一个个体,即半径。

2. 计算当前种群中个体适应度

以当前种群中的个体 $x_{i,0}$ 为半径,采用邻域粗糙集模型(FARNRS)对训练集 S 进行属性约简,采用支持向量机在约简的训练子集上进行训练,将测试集的分类精度作为个体的适应度 $f_{i,0} = f(x_{i,0})$。

3. For G=0 to T−1

(1)利用 $v_{i,G+1} = x_{i1,G} + F \cdot (x_{i2,G} - x_{i3,G})(i = 1,2,\cdots,NP)$ 进行变异操作,产生变异个体 $v_{i,G+1}$;

(2)利用 $u_{i,G+1} = \begin{cases} v_{i,G+1}, & \text{if } rand \leqslant CR \\ x_{i,G+1}, & \text{if } rand > CR \end{cases} (i = 1,2,\cdots,NP)$ 进行交叉操作,产生交叉个体 $u_{i,G+1}$;

(3)计算临时种群个体适应度 $f_{i,G+1} = f(u_{i,G+1})$;

(4) 利用贪婪准则进行选择操作,获得新一代种群 $x_{i,G+1}(i = 1,2,\cdots,NP)$;

End

4. 输出最优半径。

4 仿真实验

4.1 实验数据

为评价本文算法性能,选取 3 个标准肿瘤数据集(Leukemia[5]、Small Round Blue Cell Tumor (SRBCT)[21]、Acute Lymphoblastic Leukemia(ALL)[22])进行仿真实验,数据集描述如表 1 所列。

表 1 标准肿瘤数据集

数据集	类别	基因	样本数	训练样本	测试样本
Leukemia	3	7129	72	38	34
SRBCT	4	2308	83	63	20
ALL	6	12625	248	148	100

4.2 实验结果与分析

4.2.1 前 TOP 基因数量对分类精度的影响

本文首先利用 ReliefF 算法进行特征基因初选,需确定初选的基因子集中应该含有多少个基因才算合适。初选基因的数量往往会影响分类精度以及效率,而目前关于初选的基因子集中基因数量的确定没有定论,所以为保证本文算法的性能,通过实验来分析基因数量对分类精度的影响,进而确定初选的基因数量,结果见图 1。

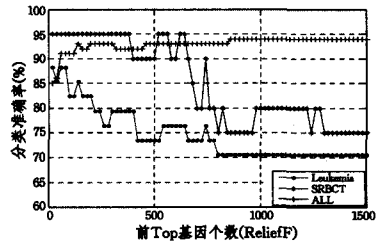

图 1 前 TOP 基因数与分类精度的关系

从图 1 清楚地看到,分类精度并非随着基因数量的增加而提高,利用 ReliefF 算法进行基因初选,前 Top 基因数量对分类精度的影响是比较大的。对于 Leukemia 和 SRBCT,其分类精度并非随着基因数量的增加而不断提高,反而当基因数量达到一定程度,其精度降低;对于 ALL,开始阶段分类精度随着基因数量的增加而提高,但基因数量增加到达一定程度后,其分类精度基本保持不变。这表明在微阵列数据中确实存在大量无关的基因,这些基因的存在既影响分类精度,还导致时间效率的降低。以上结果说明 ReliefF 方法不但能大幅降低预选基因个数,还能提高分类精度,因而是剔除与分类任务无关基因的一种有效方法。

4.2.2 邻域粗糙集半径对分类精度的影响

为了分析邻域粗糙集半径对模型性能的影响,以 0.01 为步长,将邻域粗糙集半径从 0.1 取到 1,每取一个半径,利用邻域粗糙集模型约简获得一个基因子集,共得到 100 个基因子集,并利用 RBF-SVM 进行分类。其结果见图 2—图 4。

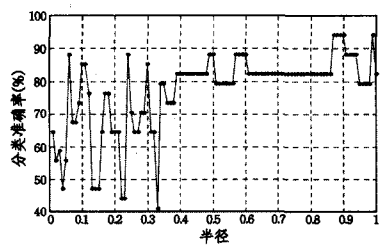

图 2 邻域半径与分类精度的关系(Leukemia)

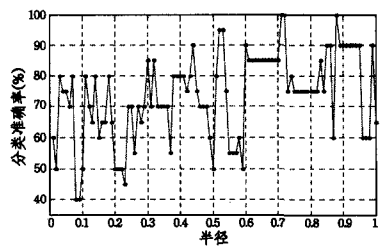

图 3 邻域半径与分类精度的关系(SRBCT)

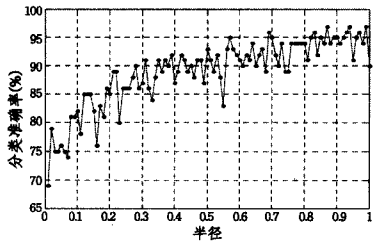

图 4　邻域半径与分类精度的关系(ALL)

从图 2—图 4 发现，邻域半径对分类精度的影响较大，不同半径对应的分类精度有着较大的差异，且半径选择得不合适不但不会提高分类精度，反而会大幅降低分类效果，这给邻域粗糙集模型的广泛使用带来局限。本文采用差分进化算法来优化半径，以获得更高、更稳定的分类性能。

4.2.3　差分进化算法的参数对其优化性能的影响

DE 算法的参数主要有种群规模、迭代次数、变异算子和交叉算子。研究[23]表明，种群规模与迭代次数相关，较大的种群规模会加快算法收敛，以减少迭代次数；反之需要较大的迭代次数才能达到收敛状态。所以种群规模往往取 10 到 30 之间，而迭代次数在 10 到 100 之间。

图 5 展示变异算子对 DE 算法优化性能的影响。变异算子 $f\in[0,2]$，决定偏差向量的放大比例。对 Leukemia：当 F 取 0 或 1.7 时，结果为 88.2%，其余取值上均获得 94.2%的最优值；对 SRBCT：当 F 取 0、0.2 或 2 时，结果为 95%，其余取值上均获得 100%的最优值；对 ALL：当 F 取 0 时，结果为 96%，其余取值上均获得 97%的最优值。总体来说，变异算子在大部分取值上均获得全局最优值，仅在个别取值上的结果不太理想，但仍然不低于其他算法的结果，所以变异算子对 DE 算法的优化性能影响不大。

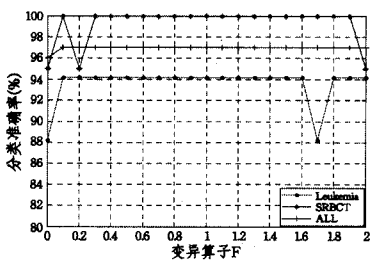

图 5　变异算子 F 对 DE 算法性能的影响

图 6 展示了交叉算子对 DE 算法优化性能的影响。交叉算子 $CR\in[0,1]$，控制试验向量参数来自于随机选择的变异向量而不是原来向量的概率。同上面分析一样，交叉算子在大部分取值上均获得全局最优值，仅在个别取值上的结果不太理想，但仍然不低于其他算法的结果，所以总的来说，交叉算子对 DE 算法的优化性能的影响也不大。

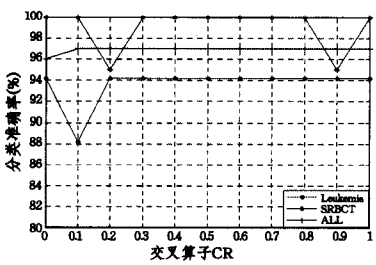

图 6　交叉算子 CR 对 DE 算法性能的影响

通过分析参数对 DE 性能的影响，本文算法中 DE 参数设定如下：种群规模 $NP=15$、变异算子 $F=1$、交叉算子 $CR=0.5$、最大迭代次数 $T=20$。

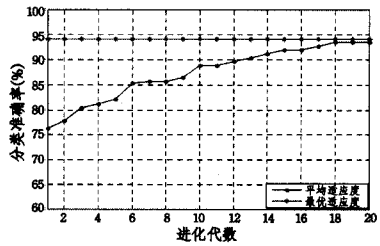

图 7　差分进化算法的半径优化迭代(Leukemia)

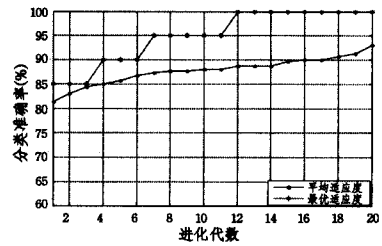

图 8　差分进化算法的半径优化迭代(SRBCT)

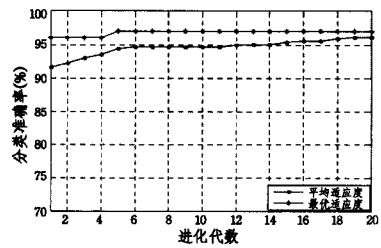

图 9　差分进化算法的半径优化迭代(ALL)

图 7—图 9 显示了利用差分进化算法优化半径的迭代过程。可清楚地看到，采用 DE 算法均能获得最优半径，对应的分类精度最高，也很稳定，而且基本不超过 20 次迭代就能获得最优半径，表明 DE 算法有着较强的全局搜索能力和搜索效率，也充分说明利用差分进化对邻域粗糙集模型进行优化是有效的、可行的。其中，Leukemia、SRBCT 和 ALL 的最优半径分别是 0.89、0.72、0.79，集中在[0.7,0.9]，当然邻域半径也与研究对象有着重要的关系，所以也不能在任何问题中均直接取较大值。

4.2.4　本文算法结果及比较

表 2 展示不同算法在 3 个数据集上的实验结果，表 3 给出本文算法所获得的特征基因。我们从两个方面分析：(1)分类精度。本文算法的分类精度最高，相比 ReliefF＋SVM 算法、SVM 算法分别至少提高了 3%和 30%。说明本文采用 ReliefF 删除大量无关基因，然后利用优化的邻域粗糙集模型约简冗余基因，使获得的特征基因子集具有更强的分类性能，更能从本质上反应数据的真实结构。(2)特征基因数量。本文算法所确定的基因个数是最少的，远远少于原基因个数以及由 ReliefF 确定的基因数量：Leukemia 从原始的 7129 个减少到 3 个，SRBCT 从原来的 2308 个减少到 5 个，ALL 从原来的 12625 个减少到 13 个。特征基因个数的减少，既可大幅提高算法的运行效率，又可增加对微阵列数据的理解程度；本文算法选择的特征基因可给临床诊疗提供更可靠的依据和参考。

(下转第 316 页)

2013,116:112-121
[13] 陈雪松,徐学军,朱洪波. 基于图像势能理论的目标轮廓特征提取方法 [J]. 计算机科学,2011,38(6):270-274
[14] 李德毅,杜鹢. 不确定性人工智能 [M]. 北京:国防工业出版社,2005
[15] 淦文燕,赫南,李德毅. 一种基于拓扑势的网络社区发现方法 [J]. 软件学报,2009,20(8):2241-2254
[16] 杨炳儒,高静,宋威. 认知物理学在数据挖掘中的应用研究 [J]. 计算机研究与发展,2006,43(8):1432-1438
[17] 王树良,邹珊珊,操保华,等. 利用数据场的表情脸识别方法 [J]. 武汉大学学报:信息科学版,2010,35(6):738-742
[18] 吴涛,金义富,侯睿,等. 不确定性边缘表示与提取的认知物理学方法 [J]. 物理学报,2013,62(6):171-183
[19] Yann Le-cun, Corinna Cortes, Christopher Burges. The MNIST database of handwritten digits [OL]. http://yann. lecun. com/exdb/mnist/

(上接第 294 页)

综合以上分析,本文算法既提高分类精度,又大大减少特征基因个数,提高算法效率,充分说明该算法能剔除微阵列数据中的无关和冗余基因,揭示数据的本质特征和结构,提高分类精度和泛化性能,该算法是一种有效的特征基因选择方法。

表 2 不同算法的实验结果

数据集	SVM		ReliefF+SVM		本文算法		
	精度	基因	精度	基因	精度	基因	半径
Leukemia	56%	7129	88.2%	100	**94.2%**	**3**	0.89
SRBCT	60%	2308	95%	1000	**100%**	**5**	0.72
ALL	68%	12625	94%	900	**97%**	**13**	0.79

表 3 本文算法选择的特征基因

数据集	特征基因
Leukemia	{U05259_rna1_at、X03934_at、L09209_s_at}
SRBCT	{770394、782811、812105、796258、207274}
ALL	{8745_at、38682_at、37643_at、40551_at、36496_at、37085_g_at、40711_at、41517_g_at、32984_s_at、37106_at、36089_at、35580_at、947_at}

结束语 本文针对微阵列数据的高维小样本特点以及高噪声和高冗余特性,提出了一种特征基因混合选择方法。ReliefF 算法剔除大量无关基因,利用优化的邻域粗糙集模型进一步清除了存在的冗余基因,所以最终获得最优特征基因子集。实验说明该算法能用较少的基因获得较大的分类精度,在提高泛化性能的同时降低了时间代价,是一种可行的、有效的特征基因选择方法。

参 考 文 献

[1] Derisi J L, Iyer V R, Brown P O. Exploring the metabolic and genetic control of gene expression on a genomics [J]. Science, 1997,278(5338):680-686
[2] Zhao Y H, Wang G R, Li Y, et al. Finding novel diagnostic gene patterns based on interesting non-redundant contrast sequence rules[C]// International Conference on Data Mining. 2011:972-981
[3] Zhao Y H, Yin Y, Wang G R. Identifying top-k vital patterns from multiclass medical data[C]// BioMedical Information Engineering. 2009:536-39
[4] 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
[5] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286:531-537
[6] Arfin S M, Long A D, Ito E T. Global gene expression profiling in esherichia coliK12: the effects of integration host factor[J]. Journal of Biological Chemistry, 2000, 275:29672-29684
[7] Tusher V G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response[J]. PNAS, 2001, 98:5116-5121
[8] Pan W. A Comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments[J]. Journal of Bioinformatics, 2002, 18:546-554
[9] Kira K, Rendell L A. A practical approach to feature selection [C]// Proceedings of the Ninth International in Machine Learning Conference. 1992:145-156
[10] Kononenko I. Estimating attributes: analysis and extensions of RELIEF[C]// Proceedings of the European Conference on Machine Learning, Lecture Notes in Computer Science. 1994, 784: 171-182
[11] 张文修,仇国芳. 粗糙集属性约简的一般理论[J]. 中国科学 E 辑:信息科学,2005,12:1304-1313
[12] Yao Y, Yao B. Covering based rough set approximations[J]. Information Sciences, 2012, 200:91-107
[13] 胡清华,于达仁. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报,2008,15(3):121-125
[14] 胡清华,赵辉,于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法[J]. 模式识别与人工智能,2008,21(6):89-95
[15] Storn R, Price K. Differential evolution simple and efficient adaptive scheme for global optimization over continuous spaces [R]. Berkeley: University of California, 2006
[16] Zhao H. Intrusion Detection Ensemble Algorithm based on Bagging and Neighborhood Rough Set[J]. International Journal of Security and Its Applications, 2013, 7(5):193-204
[17] Chen T. Classification algorithm on gene expression profiles of tumor using neighborhood rough set and support vector machine [J]. Advanced Materials Research, 2014, 850:1238-1242
[18] 赵晖. 融合邻域粗糙集与粒子群优化的网络入侵检测[J]. 计算机工程与应用,2013,18:73-77
[19] 赵晖. 基于邻域粗糙集与 KNN 的网络入侵检测[J]. 河南科学,2013,9:1404-1408
[20] 雍龙泉. 求解一类多目标优化问题的极大熵差分进化算法[J]. 中南大学学报,2013,S2:160-164
[21] Khan J. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. Nature Medicine, 2001, 7(6):673-679
[22] Yeoh E J. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling[J]. Cancer Cell, 2002, 1(2):133-143
[23] 周艳平,顾幸生. 差分进化算法研究进展[J]. 化工自动化及仪表,2007,34(3):1-5