基于 ReliefF 和蚁群算法的特征基因选择方法*

吴辰文，李晨阳，郭叔瑾，闫光辉
（兰州交通大学 电子与信息工程学院，兰州 730070）

摘 要：针对高维小样本的 DNA 微阵列数据多分类问题，提出一种基于 ReliefF 和蚁群算法的特征基因选择方法（ReliefF and ant colony optimization，ReFACO）。该方法首先采用 ReliefF 算法评估特征权重，根据阈值筛选出无关基因；然后引入改进的蚁群算法，在迭代改进的过程中寻找最优基因子集；最后利用经典分类算法对维数约简后的数据分类识别。经实验证明，该方法可有效地剔除无关和冗余基因，并利用较少特征基因达到较高多分类效果。

关键词：DNA 微阵列数据；ReliefF 算法；蚁群算法；特征选择
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2018)09-2610-04
doi:10.3969/j.issn.1001-3695.2018.09.011

Feature gene selection method based on ReliefF and ant colony optimization

Wu Chenwen, Li Chenyang, Guo Shujin, Yan Guanghui
(*School of Electronics & Information Engineering, Lanzhou Jiaotong University, Lanzhou* 730070, *China*)

Abstract: Aiming at the multi-classification problem of DNA microarray data with the characteristic of high dimension and small sample, this paper proposed a feature gene selection algorithm based on ReliefF and ant colony optimization (ReFACO). The method adopted ReliefF algorithm to evaluate the feature weights, and selected the irrelevant genes based on the threshold, introduced an improved ant colony algorithm to find the optimal subset of genes in the process of iteration and improvement, used classical classification algorithms to classify and identify the data set, dimensions of which had been reduced. Experimental results show that the method can eliminate irrelevant and redundant genes effectively, and achieve a higher classification performance with less characteristic genes.

Key words: DNA microarray data; ReliefF algorithm; ant colony optimization; feature selection

0 引言

微阵列技术是一种快速发展的分子生物学技术，能够在单一的实验中检测数以万计的基因。目前，微阵列技术的应用非常广泛，包括疾病监测、环境监测、进化生物学、细胞生理学、药物开发与生产[1~3]等。微阵列数据的分类是通过利用机器学习和数据挖掘技术，建立一个模型来对样本进行分类，从而提取有价值的信息。DNA 微阵列数据的高维、小样本特点，导致了“维数灾难”问题的出现。研究表明，在数量庞大的 DNA 微阵列数据中，参与分类的有效基因只占总体的很少部分，大多数基因不包含与分类任务相关的信息。因此，数据预处理是实现基因表达数据分类高效、准确、可靠的必要步骤。

特征基因选择方法是从原始基因组中确定一组最有效的基因子集，从而减少计算成本，提高分类精度。目前所提出来的特征选择方法可以分为过滤式（filter）、封装式（wrapper）、嵌入式（embedded）三类[4,5]。过滤式方法利用数据自身的统计特性作为基因的评价准则来选择特征子集，其时间复杂度低，但分类准确率也较低。封装式方法是在基因选择过程中应用特定的学习模型来评估所选择的特征子集，并依靠学习模型的准确性来指导搜索过程。由于封装式包含给定的学习模型，并且考虑基因之间的相互关系，所以准确率高于过滤式方法，但是其时间复杂度远高于过滤式方法。嵌入式方法利用初始基因数据训练学习模型，并在该过程中完成基因的选择。嵌入式方法的主要优点在于与学习模型相互作用，但其时间复杂度同样较高。

蚁群算法是一种用于求解组合优化问题的元启发式（MetaHeuristic）方法，具有健壮性强、回馈控制和多 agent 系统等特点[6]，算法的贪心策略和随机策略增加了其全局搜索能力。在大多数的微阵列数据中，存在被错误标记或其类别标签不可靠的样品，蚁群算法可通过特征之间的冗余性分析，避免了监督学习算法中存在的过度依赖标签的情况。目前，基于蚁群算法的特征选择方法得到了广泛关注，例如，文献[7]介绍了一种基于蚁群算法的无监督概率特征选择方法；文献[8]提出了一种改进的蚁群优化特征选择算法，并通过模糊逻辑控制系统来调整算法中的动态和静态参数；文献[9]将蚁群优化算法与粗糙集相结合，提出一种以互信息作为启发式信息的特征选择方法。

Relief 系列算法是一种典型的 Filter 方法。作为众所周知的效率较高的维数约简方法，Relief 系列算法主要包括最初由 Kira 和 Rendell 提出的 Relief 和后来扩展的 Relief 与 ReliefF，其中 ReliefF 主要用于解决多分类、数据缺失和存在噪声等问题[10]。Relief 系列算法时间复杂度低，不使用分类精度作为评价函数，但由于它是基于特征权重的算法，在进行特征选择时仅提高与标签关联度高的特征权重值，剔除权重值低的特征，所以不能有效地去除冗余特征。

因此，考虑 Relief 系列算法的计算效率和蚁群算法的良好性能，本文提出一种基于 ReliefF 和蚁群算法的特征基因选择算法。首先利用 ReliefF 算法去除权重较低的特征；然后采用蚁群算法在迭代改进的过程中进行特征选择，并引入本文提出

收稿日期：2017-04-18；**修回日期**：2017-06-02 **基金项目**：国家自然科学基金资助项目(61163010)；甘肃省自然科学基金资助项目(1308RJZA111)

作者简介：吴辰文(1964-)，男，甘肃靖远人，教授，硕士，主要研究方向为数据挖掘、可视化、网络安全(wuchenwen@mail.lzjtu.cn)；李晨阳(1991-)，男，河南开封人，硕士研究生，主要研究方向为数据挖掘；郭叔瑾(1992-)，女，甘肃定西人，硕士研究生，主要研究方向为数据挖掘、可视化；闫光辉(1970-)，男，河南商丘人，教授，博士，主要研究方向为数据挖掘、数据仓库等.

的适应度函数来评估所选择的基因子集;最后在几种常见的基因表达数据集上进行实验,其中包括两个二分类数据集和两个多分类数据集。

1　ReliefF 算法

Relief 算法作为一种高效率的 Filter 算法,它根据特征重要性进行次序排列,并将高于指定阈值的特征作为特征子集。Relief 从训练集中任意选取样本 R,对于每个 R 有两个最近邻:一个来自同类样本 H,另一个来自异类样本 M。若样本 R 与 H 关于特征 A 在训练集中存在差异,则特征 A 将被赋予较低权重;而样本 R 与 M 关于特征 A 在训练集中存在差异,则特征 A 将被赋予较高权重。Relief 由式(1)更新特征 A 的权重 $W[A]$。

$$W[A]=W[A]-\mathrm{diff}(A,R,H)/m+\mathrm{diff}(A,R,M)/m \tag{1}$$

其中:m 是随机抽样个数;diff 函数表示给定属性的两个样本之间的差异。在计算 $W[A]$ 时,用 m 进行归一化处理,保证权重值在 $-1\sim1$。

当特征属性为标称属性时,$\mathrm{diff}(A,I_x,I_y)$ 计算[11]如下:

$$\mathrm{diff}(A,I_x,I_y)=\begin{cases}0 & \mathrm{value}(A,I_x)=\mathrm{value}(A,I_y)\\ 1 & \text{otherwise}\end{cases} \tag{2}$$

当特征属性为数值属性时,$\mathrm{diff}(A,I_x,I_y)$ 计算如下:

$$\mathrm{diff}(A,I_x,I_y)=\frac{|\mathrm{value}(A,I_x)-\mathrm{value}(A,I_y)|}{\max(A)-\min(A)} \tag{3}$$

传统的 Relief 算法只局限于二分类问题,并且不能处理数据的缺失和存在噪声等问题。1994 年 Kononenko 扩展了 Relief 算法,提出 ReliefF 算法并解决了上述问题。ReliefF 通过平均每个样本同类的 k 个最近邻差异值和其他异类的 k 个最近邻差异值来降低数据中噪声的影响。最近邻样本数的选择是 Relief 与 ReliefF 之间最基本的差异,其确保了 ReliefF 算法的鲁棒性。为了处理数据的缺失问题,Kononenko[12]对 $\mathrm{diff}(A,I_x,I_y)$ 函数进行改进,使得属性的缺失值可由概率计算得出。

若样本 I_x 有缺失值,则 $\mathrm{diff}(A,I_x,I_y)$ 计算如下:

$$\mathrm{diff}(A,I_x,I_y)=1-P(\mathrm{value}(A,I_y)\mid\mathrm{class}(I_x)) \tag{4}$$

若样本 I_x 和 I_y 都有缺失值,则 $\mathrm{diff}(A,I_x,I_y)$ 计算如下:

$$\mathrm{diff}(A,I_x,I_y)=1-\sum_{V}^{\#\mathrm{value}(A)}(P(V\mid\mathrm{class}(I_x))\times P(V\mid\mathrm{class}(I_y))) \tag{5}$$

其中:V 表示所有样本中特征 A 的值。ReliefF 算法描述如算法 1 所示。

算法 1

输入:训练集 D,迭代次数 m,最近邻样本数 k。
输出:特征权重向量 W。
初始化特征权重;
　for $i=1$ to m
　　从 D 中随机选择一个样本 R_i;
　for each class $C=\mathrm{class}(R_i)$
　　找到与 R_i 同类的 k 个最近邻样本 H_j;
　for each class $C\neq\mathrm{class}(R_i)$
　　找到与 R_i 不同类的 k 个最近邻样本 $M_j(C)$;
　for $A=1$ to all feature

$$W[A]=\frac{W[A]\sum_{j=1}^{k}\mathrm{diff}(A,R_i,H_j)}{mk}+\sum_{C\neq\mathrm{class}(R_i)}\frac{p_i(C)}{1-p(\mathrm{class}(R_i))}\times\sum_{j=1}^{k}\frac{\mathrm{diff}(A,R_i,M_j(C))}{mk}$$

ReliefF 每次随机从 D 中任意选取样本 R_i,然后从相同类别的样本中找出 k 个最近邻样本 H_j,从每个与 R_i 不同类的样本中找出 k 个最近邻样本 $M_j(C)$,随后更新权重,重复以上过程 m 次。最后对特征权重排序,筛选出权重较低的特征。

2　基于蚁群算法的特征选择方法

蚁群算法是一种生物智能算法,由 Dorigo 等人于 1991 年提出,以解决 TSP 问题[13]。蚂蚁搜寻食物时会在地面上留下信息素,信息素强度取决于蚁穴与食物之间的距离,较短的路径留下较多信息素。当新蚂蚁进入系统时,喜欢沿着有更多信息素的边搜寻;之后,由于正反馈效果,其余蚂蚁也都会采选较短路径。鉴于蚁群算法的健壮性强、回馈控制和多 agent 系统等特点,本章将提出一种基于蚁群算法的特征选择方法。该方法可以有效地解决多分类问题,并剔除数据集中的冗余特征。

2.1　搜索空间

所提出方法的搜索空间为完全图,图中节点代表原始特征集合,特征 v_i 与 $v_j(\forall i,j=1,2,\cdots,n)$ 之间边的权重值为两者的相似度值,定义如下:

$$\mathrm{sim}(v_i,v_j)=\left|\frac{v_iv_j}{\|v_i\|\,\|v_j\|}\right|=\left|\frac{\sum_{s=1}^{p}v_{si}v_{sj}}{\sqrt{\sum_{s=1}^{p}v_{si}^2}\sqrt{\sum_{s=1}^{p}v_{sj}^2}}\right| \tag{6}$$

其中:p 是样本数量;v_{si} 和 v_{sj} 为样本 s 中第 i 个和第 j 个特征。

此外,特征之间相似度值的倒数作为启发式信息可以指导蚂蚁向最优结果搜寻,信息素值 $\tau_{ij}(\forall i,j=1,2,\cdots,n)$ 的初始强度在算法开始搜索之前被设为常数,并在搜索过程中由蚂蚁更新,信息素反映了蚂蚁从过去经验获得的信息。

2.2　状态转移规则

在所提出算法中,信息素强度和启发式信息对是否能够找到最优子集起到至关重要的作用。状态转移规则由贪心策略和随机策略组成,蚂蚁通过该规则选择下一个特征。

贪心策略中,在特征 v_i 上的第 k 个蚂蚁由式(7)选定特征 v_j。

$$v_j=\arg\max_{u\in J_i^k}\{[\tau_{iu}][\eta(v_i,v_u)]^\beta\}\quad \text{if } q\leqslant q_0 \tag{7}$$

随机策略中,第 k 个蚂蚁通过概率 $P_k(v_i,v_j)$ 选择下一个特征 v_j,定义[14]如下:

$$P_k(v_i,v_j)=\begin{cases}\dfrac{[\tau_{ij}][\eta(v_i,v_j)]^\beta}{\sum_{u\in j_i^k}[\tau_{iu}][\eta(v_i,v_u)]^\beta} & \text{if } v_j\in J_i^k\\ 0 & \text{otherwise}\end{cases}\quad \text{if } q>q_0 \tag{8}$$

其中:J_i^k 是未被访问的特征集合;τ_{iu} 是边 (v_i,v_u) 的信息素值;$\eta(v_i,v_u)=\dfrac{1}{\mathrm{sim}(v_i,v_u)}$ 是启发式信息;参数 $\beta(>0)$ 控制信息素浓度与启发式信息的权重关系;$q_0\in[0,1]$ 为调控贪心策略和随机策略发生概率的参数,随机变量 q 在 $[0,1]$ 上服从均匀分布。

结合贪心策略和随机策略的状态转移规则,可以有效地避免蚂蚁在选择特征子集时陷入局部最优解的问题。贪心策略增强了蚂蚁的局部搜索能力,而随机策略则提供了蚂蚁寻找更多解的可能,扩大了其解空间。

2.3　信息素更新规则

当全部蚂蚁都搜索结束之后,图中信息素开始更新。信息素值更新规则如下:

$$\tau_{ij}(t+1)=(1-\rho)\tau_{ij}(t)+\frac{EC[i,j]}{\sum_{u,v=1,2,\cdots,n}EC[u,v]}+\sum_{k=1}^{A}\Delta\tau_{ij}^k(t) \tag{9}$$

$$\Delta\tau_{ij}^k(t)=\begin{cases}\mathrm{fitness}(k) & \text{if } (i,j)\in\mathrm{subset}(k)\\ 0 & \text{otherwise}\end{cases} \tag{10}$$

其中:$\rho\in(0,1)$ 是信息素蒸发速率;$\tau_{ij}(t)$ 和 $\tau_{ij}(t+1)$ 分别表示在 t 和 $t+1$ 时刻边 (v_i,v_j) 上的信息素值;$\Delta\tau_{ij}^k(t)$ 是第 k 只蚂蚁在 t 时刻经过边 (v_i,v_j) 时释放的信息素值;$EC[i,j]$ 表示蚂蚁访问边 (v_i,v_j) 的次数;A 是蚂蚁的数量。蚂蚁 k 选择的特征子集 $\mathrm{subset}(k)$ 的适应度函数计算如下:

$$\mathrm{fitness}(k)=\frac{1}{|\mathrm{subset}(k)|}\sum_{i=1}^{|\mathrm{subset}(k)|}\frac{1}{LS(v_i)} \tag{11}$$

其中:$|\mathrm{subset}(k)|$ 表示子集 k 中的特征数量;v_i 是 $\mathrm{subset}(k)$ 中的第 i 个特征;$LS(v_i)$ 是特征 v_i 的拉普拉斯分值。

拉普拉斯分值[15,16]是由 He 等人基于 Laplacian Eigenmaps 和局部保留投影而提出的,其主要思想是将数据集中的样本点

映射到图中的节点,并构造最近邻图 G,将最近邻节点之间的相似度作为权重值。拉普拉斯分值定义如下:

$$LS(v_i) = \frac{\sum_{m,n}(v_{mi} - v_{ni})^2 S_{mn}}{\mathrm{var}(v_i)} \tag{12}$$

其中:$v_{mi} - v_{ni}$表示特征 v_i 中样本 m 与 n 的差值;$\mathrm{var}(v_i)$表示特征 v_i 的方差;S_{mn}是样本 m 和 n 的相似度,其计算如下:

$$S_{mn} = \begin{cases} e^{-\frac{\|m-n\|^2}{t}} & \text{if } m,n \text{ is connected} \\ 0 & \text{else} \end{cases} \tag{13}$$

其中:t 为常量。特征的拉普拉斯分值越小,则证明该特征对局部结构保持能力越强。在信息素更新规则中存在两种策略:a)信息素蒸发策略,之前蚂蚁沉积的信息素强度随时间降低,其目的是避免算法局部收敛过快;b)信息素沉积策略,在搜索过程中,较优路径会沉积更多的信息素,以便蚂蚁在进一步的迭代中探索搜索空间中的最佳区域。本节所提出的适应度函数可以对特征的相关性进行评估,进而筛选掉其中的冗余特征。

ACO 算法的框架如算法 2 所示。

算法 2

输入:训练集 $D(p \times n$ 矩阵),迭代次数 I,蚂蚁个数 A,每次迭代中蚂蚁选择的特征数量 NG。

输出:约简后的训练集 $\tilde{D}$。

```
计算特征之间的相似度 sim(v_i,v_j), ∀i,j=1,2,…,n;
信息素 τ_ij(1)的初始强度设为常数, ∀i,j=1,2,…,n;
for t=1 to I
    将边计数器 EC[i,j]初始化为零, ∀i,j=1,2,…,n;
    将蚂蚁任意地放在图中的节点;
      for k=1 to A
      for i=1 to NG
        根据状态转换规则选择下一个特征;
        将第 k 个蚂蚁移至新选择的特征,增加访问边对应的边计数器;
        使用适应度函数评估所选特征子集;
        寻找全局最优解;
        应用信息素更新规则更新信息素值;
  在迭代中保持全局最佳子集;
  输出全局最佳子集。
```

所提出的方法由初始化和迭代选择部分组成。在初始化部分中,计算了基因之间的相似度值作为启发式信息,并对信息素进行初始化。迭代选择部分是一个在迭代中持续改进的过程,在每次迭代中,首先将蚂蚁随机放置图中,根据状态转移规则选择候选子集,对候选子集进行评估后保留最佳子集,最后根据信息素更新规则更新信息素。当迭代 I 次后,输出全局最佳子集。

3 基于 ReliefF 与蚁群算法的特征基因选择方法(ReFACO)

目前,大多数 DNA 微阵列数据除了高维、小样本的特点之外,还存在多分类、数据缺失和存在噪声等问题,这使得人们在进行特征选择和分类时更具有挑战性。

本文提出一种基于 ReliefF 与蚁群算法的特征基因选择方法,该方法主要通过分析基因之间的相关性和冗余性来决定基因子集。由于 ReliefF 算法为监督学习算法,而蚁群算法为无监督学习算法,所以本文所提出的 ReFACO 算法为半监督特征选择算法。首先采用 ReliefF 算法根据基因对近距离样本的区别能力来评估该基因,通过阈值的设定,筛选出与类别标签相关性较强的基因作为候选基因子集;然后将候选基因子集输入蚁群算法,在蚁群算法的迭代改进过程中选择最佳基因子集。算法流程如图 1 所示。

ReliefF 作为一种可以处理多分类问题的算法,它能够根据基因与类别的依赖关系,较快速地提供基因权重评估,有效地剔除无关基因;基于群体智能的蚁群算法具有良好的健壮性和分布式计算能力,状态转移规则中随机策略和贪心策略提高了全局搜索的性能,在信息素更新规则中以拉普拉斯分值作为适应度函数对每个基因进行评估,确保了所选基因子集为全局最优子集。

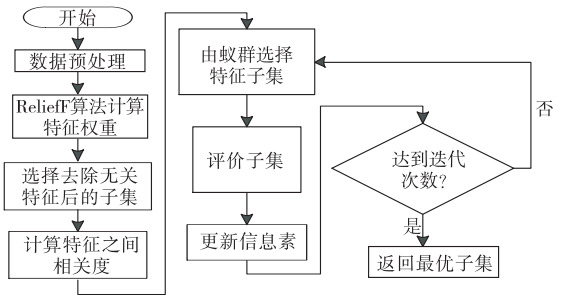

图 1 算法流程

4 实验结果分析

为了评估 ReFACO 算法对 DNA 微阵列数据的处理性能,将其使用在结肠癌、SRBCT、白血病和肺癌四个基因数据集中。以上四个数据集都是目前研究使用较多且公开的标准数据集。数据集的简要描述如表 1 所示。实验平台为 PC 机(Windows 7,Intel Core i5-2410 M CPU@2.30 GHz),使用的软件为 MATLAB 2014a 和 R。本文所使用的经典分类器是 SVM 和 C4.5。其中,SVM 采用一对多策略[17]来解决多分类问题,它在解决高维小样本问题方面有独特的优势;C4.5 算法[18]可以对缺失数据进行处理。

表 1 基因数据集

数据集	基因	样本	类别
结肠癌	2 000	62	Tumor(40)、Normal(22)
SRBCT	2 308	83	EWS(29)、BL(11) NB(18)、RMS(25)
白血病	7 129	72	ALL-T(9)、ALL-B(38) AML(25)
肺癌	12 600	203	Adeno(139)、NORM(17) Squamous(21)、COID(20)、SMCL(6)

实验主要分为两部分:a)分析比较由 ReFACO 所选择的不同特征数对分类精度的影响;b)不同特征选择方法的准确率比较。在进行实验之前,采用 Min-Max 标准化方法对所有的数据集进行预处理,将其归一化在[0,1]区间。在所提出的 ReFACO 算法中,需对参数进行设置,其中近邻值 k 为 10,蚂蚁个数为 100,参数 β 设置为 1,信息素蒸发速率 ρ 设置为 0.2。另外,采用十折交叉验证法计算分类的准确率。

4.1 特征数对分类精度的影响

为了探究 ReFACO 算法所选特征数对分类精度的影响,本组实验采用 SVM 和 C4.5 分类算法对不同数目的所选基因进行测试。图 2 和 3 表示在四种数据集上运行所提出方法,并使用 SVM 和 C4.5 评估不同基因数的分类准确率。

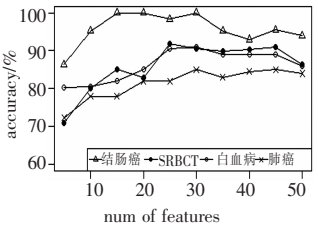

图2 SVM 对不同数目特征的分类精度

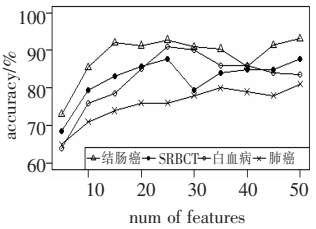

图3 C4.5 对不同数目特征的分类精度

从图 2、3 的实验结果可以看出,在使用 SVM 和 C4.5 算法对四种数据分类时,随着特征维数的增多,分类准确率总体呈上升趋势,其中采用 SVM 进行分类时,结肠癌、SRBCT、白血病和肺癌分别在特征个数为 15、25、25、30 时达到最好的分类效果;而采用 C4.5 进行分类时,以上四种数据分别在特征个数为 15、25、25、35 时达到最好的分类效果。这说明随着特征子

集中元素的增多,原数据集中的部分细节与特征之间隐含的联系也逐渐增多;但是当特征增多到一定程度后,原数据集中的无关特征和噪声会降低分类的准确率。另外,对比图 2、3 发现,在对 DNA 微阵列数据进行分类时,SVM 算法的准确率会明显高于 C4.5 算法,这说明 SVM 算法在处理高维小样本数据时有更强的泛化能力。

4.2　特征选择方法的比较

为了充分证明所提出算法的有效性,本组实验选择四种经常使用的特征基因选择方法与所提出方法进行比较。其中,FCBF(fast correlation based filter)[19]是一种基于互信息的快速滤波算法,其利用对称的不确定度量来评估特征之间的相关性;CFS(correlation-based feature selection)[20]采用基于特征与标签相关性的启发式方法来评价特征的重要性。mRMR-ReliefF[21]是一种二阶消除算法,该算法首先应用 ReliefF 筛选掉权重较低的特征,然后通过冗余度和相关性计算来选择特征子集。

表 2 给出了五种特征选择算法在四种数据集上所选特征个数。其中,FCBF 算法所选择的特征个数最少,ReFACO 算法次之,而 mRMR-ReliefF 算法所选特征最多。图 4、5 展示了五种算法分别基于 SVM 和 C4.5 的分类正确率。从该实验结果可以发现,在所应用的四种数据集上,本文提出的 ReFACO 算法所选特征的分类准确率总体较高,mRMR-ReliefF 算法次之,而 ReliefF 算法最差。采用 SVM 算法对结肠癌数据进行分类时,mRMR-ReliefF 和 ReFACO 算法都可以达到完全正确,在对另外三种数据进行分类时,ReFACO 算法的准确率显然高于另外四种。C4.5 算法的分类结果略有不同,在对白血病数据进行分类时,ReFACO 算法的结果并不理想,正确率与 mRMR-ReliefF 算法相比略有下降,而另外三种数据的分类准确率依然是 ReFACO 算法较高。从表 1 可知,在本实验所采用的数据中,结肠癌与白血病都是二分类数据,而 SRBCT 与肺癌则是多分类数据。因此,在处理二分类问题时,mRMR-ReliefF 和 ReFACO 算法的性能较为优秀,而 ReliefF、FCBF 和 CFS 算法都远不及前两者;在处理多分类问题时,ReFACO 算法显然优于另外几种。这是因为 ReFACO 算法是一种基于 Filter 算法和群体智能的方法,在其无关基因筛选阶段,通过差异值计算,有效地避免了数据的缺失问题;在其冗余特征筛选阶段,随机探索搜索空间,并在迭代改进的过程中选择特征子集,最终使算法不但适合多分类问题,而且相比其他算法有更高的准确率。结合表 2 可以得出,ReFACO 算法可以在选择较少特征的情况下实现最高效率的维数约简。

表 2　五种特征选择算法所选特征个数

数据集	ReliefF	mRMR-ReliefF	CFS	ReFACO	FCBF
结肠癌	17	23	22	15	9
SRBCT	21	26	26	25	17
白血病	28	31	34	25	12
肺癌	32	45	32	30	21
平均值	25	31	29	24	15

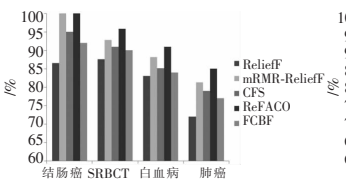

图4　五种算法基于 SVM 的分类性能比较

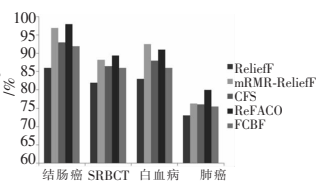

图5　五种算法基于 C4.5 的分类性能比较

5　结束语

本文提出一种基于 ReliefF 和蚁群算法的半监督特征基因选择方法,该算法主要针对目前 DNA 微阵列数据维数高、样本少的特点,通过结合 Filter 方法的计算效率和蚁群算法出色的搜索能力来提高算法的特征选择性能。为了研究 ReFACO 算法对 DNA 微阵列数据的处理能力,本文将其分别应用在两个二分类基因数据集和两个多分类基因数据集,并采用 SVM 和 C4.5 算法对所选基因子集进行评估。实验结果表明,ReFACO 算法不但可以将无关基因和冗余基因去除,而且能提升多分类模型的准确度,从而证明该方法的可行性和有效性。后续的工作将侧重于研究 ReFACO 算法中参数的动态调整方法,实现对算法的进一步优化。

参考文献:

[1] 刘曦,刘卓琦,罗达亚. 基因表达谱微阵列网络数据库在肿瘤研究中的应用[J]. 中国生物化学与分子生物学报,2016,32(3):260-266.

[2] Butt H Z,Sylvius N,Salem M K,*et al*. Microarray-based gene expression profiling of abdominal aortic aneurysm[J]. European Journal of Vascular & Endovascular Surgery the Official Journal of the European Society for Vascular Surgery,2016,52(1):47-55.

[3] Hsieh S Y,Chou Yuchun. A faster cDNA microarray gene expression data classifier for diagnosing diseases[J]. IEEE/ACM Trans on Computational Biology & Bioinformatics,2016,13(1):43-54.

[4] 姚旭,王晓丹,张玉玺,等. 特征选择方法综述[J]. 控制与决策,2012,27(2):161-166.

[5] Lorena L H,Carvalho A,Lorena A C. Filter feature selection for one-class classification[J]. Journal of Intelligent & Robotic Systems,2015,80(1):227-243.

[6] Yang Qiang,Chen Weineng,Yu Zhengtao,*et al*. Adaptive multimodal continuous ant colony optimization[J]. IEEE Trans on Evolutionary Computation,2017,21(2):191-205.

[7] Dadaneh B Z,Markid H Y,Zakerolhosseini A. Unsupervised probabilistic feature selection using ant colony optimization[J]. Expert Systems with Applications an International Journal,2016,53(7):27-42.

[8] Wang Gang,Chu H E,Zhang Yuxuan,*et al*. Multiple parameter control for ant colony optimization applied to feature selection problem[J]. Neural Computing and Applications,2015,26(7):1693-1708.

[9] Chen Yumin,Miao Duoqian,Wang Ruizhi. A rough set approach to feature selection based on ant colony optimization[J]. Pattern Recognition Letters,2010,31(3):226-233.

[10] Reyes O,Morell C,Ventura S. Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context[J]. Neurocomputing,2015,161(8):168-182.

[11] Kira K,Rendell L A. A practical approach to feature selection[C]// Proc of the 9th International Workshop on Machine Learning. San Francisco:Morgan Kaufmann Publishers Inc,1992:249-256.

[12] Kononenko I. Estimating attributes:analysis and extensions of RELIEF[M]//Machine Learning. Berlin:Springer,1994:356-361.

[13] Yang Jianyi,Ding Ruifeng,Zhang Yuan,*et al*. An improved ant colony optimization (I-ACO) method for the quasi travelling salesman problem Quasi-TSP[J]. International Journal of Geographical Information Science,2015,29(9):1534-1551.

[14] Colorni A,Dorigo M,Maniezzo V. Distributed optimization by ant colonies[C]//Proc of the 1st European Conference on Artificial Life. 1991:134-142.

[15] 洪小娇,彭淑娟,柳欣. 基于拉普拉斯分值特征选择的运动捕获数据关键帧提取[J]. 计算机工程与科学,2015,37(2):365-371.

[16] He Xiaofei,Cai Deng,Niyogi P. Laplacian score for feature selection[C]//Proc of International Conference on Neural Information Processing Systems. Canbridge:MIT Press,2005:507-514.

[17] 常甜甜. 支持向量机学习算法若干问题的研究[D]. 西安:西安电子科技大学,2010.

[18] 苗煜飞,张霄宏. 决策树 C4.5 算法的优化与应用[J]. 计算机工程与应用,2015,51(13):255-258.

[19] Hall M A. Correlation-based feature selection for discrete and numeric class machine learning[C]//Proc of the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc,2000:359-366.

[20] Yu Lei,Liu Huan. Feature selection for high-dimensional data:a fast correlation-based filter solution[C]//Proc of the 20th International Conference Machine Learning. Washington DC:IEEE Computer Society,2003:856-863.

[21] Zhang Yi,Ding C,Li Tao. Gene selection algorithm by combining reliefF and mRMR[J]. BMC Genomics,2008,9(2):1-10.