doi:10.3969/j.issn.1001-4616.2024.01.013

基于 F-score 和二进制灰狼优化的肿瘤基因选择方法

穆晓霞,郑李婧

(河南师范大学计算机与信息工程学院,河南 新乡 453007)

[**摘要**] 针对肿瘤基因数据维度高、噪声多、冗余性高的现状,结合 Spearman 相关系数改进 F-score 算法,在此基础上优化二进制灰狼算法,提出了一种基于改进 F-score 和二进制灰狼算法的肿瘤基因选择算法. 首先,考虑特征之间的相关性,计算每个特征的 F-score 值和特征之间的 Spearman 相关系数的绝对值;然后,计算权重系数得出各个特征的权重值,依据重要性进行排序,选出初选特征子集;最后,通过收敛因子的衰减曲线和初始化方法优化二进制灰狼算法,调整全局搜索和局部搜索所占比例,增强全局搜索能力并提高局部搜索速度,有效节省时间开销,提升特征选择的分类性能和效率,得到最优特征子集. 在 9 个肿瘤基因数据集上测试所提算法,在分类准确率和筛选特征数目两个指标上进行仿真实验,并与 4 种其他算法进行对比,实验结果证明所提算法表现良好,可有效降低基因数据维度,并具有较好的分类精度.

[**关键词**] 肿瘤基因,Fisher-score,Spearman 相关系数,二进制灰狼优化算法,特征选择
[**中图分类号**]TP311　[**文献标志码**]A　[**文章编号**]1001-4616(2024)01-0111-10

Tumor Gene Selection Based on F-score and Binary Grey Wolf Optimization

Mu Xiaoxia, Zheng Lijing

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

Abstract: According to the tumor gene situation of high dimensionality, noise and redundancy, this paper improved the F-score algorithm by the Spearman correlation coefficient, optimized the binary gray wolf algorithm, and proposed a gene feature selection algorithm with the improved F-score and the binary gray wolf algorithm. Firstly, by considering the correlation between features, the F-score value of each feature and the absolute value of Spearman correlation coefficient between features were calculated. Secondly, by calculating the weight coefficient, the weight value of each feature was derived to be ranked according to their importance and select a primary feature subset. Finally, the binary gray wolf algorithm was optimized through adjusting the proportion of global search and local search to enhance the global search capability and improve the speed of local search, so that the time overhead could be saved and the optimal feature subset was selected, which can improve the classification performance and efficiency of feature selection. The designed algorithm is tested on nine tumor gene datasets and simulated on two indexes of correct accuracy and number of filtered features. When compared with four other algorithms, the experimental results prove that the algorithm performed well, reduced the dimensionality of gene data, and had better classification accuracy.

Key words: tumor gene, Fisher-score, Spearman correlation coefficient, binary grey wolf optimization algorithm, feature selection

随着基因芯片技术的发展,研究人员能够快速、方便地获取大量基因表达谱数据[1]. 这些典型的高维数据通常拥有成千上万个特征,其中包含了大量与疾病诊断无关、被噪音污染、冗余的特征. 如何从基因表达谱数据中挖掘有价值的特征,方便研究人员研究新的药品对症下药,受到越来越多研究者的关注[2]. 特征选择作为模式识别和机器学习中关键的数据预处理步骤,是从原始特征中选择出最有效的特

收稿日期:2023-05-31.
基金项目:国家自然科学基金项目(61772176).
通讯作者:穆晓霞,博士,副教授,研究方向:机器学习、数学建模等. E-mail:muxx1980@126.com

征降低数据集维度的过程. 依据是否独立于后续的学习算法,可分为过滤式、封装式和嵌入式[3]. 过滤式策略与后续学习算法无关,一般直接利用所有训练数据的统计性能评估特征,速度快[3],如 F-score 算法、Relief 算法等,但评估与后续学习算法的性能偏差较大. 封装式策略利用后续学习算法的训练准确率评估特征子集,偏差小,但计算量大;元启发式算法作为封装式策略已广泛用于优化特征选择问题[4],如灰狼优化算法、遗传算法、粒子群优化算法等. 嵌入式算法使用带惩罚项的基模型进行特征选择,是对特征打分的预选模型,具有效果好、速度快且分类效果明显的优点,但是参数设置的是否得当需要多次调试,如随机森林、决策树算法等.

F-score 是一种简单快捷,衡量两类数据的特征选择模型[5],但无法满足多分类的实际需求. 谢娟英等[6-7]提出了基于 F-score 与支持向量机(SVM)的特征选择方法,将二分类转换为多分类,确保度量样本在多类之间辨别分析. 该作者还结合 F-score 和核极限学习机设计了一种集成特征选择算法,极大程度确保筛选出贡献程度大的基因,但是仅在 F-score 结果后融入新算法. 吴晓燕等[8]使用类间散度度量和最大互信息系数对 F-score 算法进行改进,增强了在分布不均匀但属于同一个类别中心时候的筛选情况,但是需要针对不同情况分别修正系数,增加算法的复杂度. 秦喜文等[9]引入了特征排序系数,在 F-score 没有考虑到特征间相互影响的不足上进行改进,揭示了特征在不同标签数据集下的重要性,但是需要穷举测试,耗费时间. 本文针对上述问题,引入 Spearman 系数考虑特征之间的相似性,提出了改进的 F-score 特征选择算法,清除冗余基因.

目前,灰狼优化算法[10]具有较强的收敛性、参数少、易实现等特点而广受关注. Emary 等[11]将灰狼算法修改为二进制灰狼优化算法(binary grey wolf optimization algorithm,BGWO),并运用于解决特征选择优化问题上. 王琛等[4]改进了二进制灰狼算法的适应度函数,进行特征子集合并与交叉操作,筛选最优特征子集,但是该算法的内部参数还需进一步优化. 陈长倩等[12]引入高斯分布虚线对种群初始化,通过概率实施新的二进制转换函数,优化了 BGWO 算法,但是收敛时间长. 邢燕祯等[13]通过调整全局搜索与局部搜索所占比例,提出了新型收敛因子,优化了收敛速度,但是收敛因子是在分情况讨论下的线性变化. 本文针对上述问题,提出了初始化种群策略和新的收敛因子确定方法,采用改进的 F-score 算法和 Spearman 相关系数计算特征的选择概率,进行特征选择,有效降低模型的时间开销. 同时,调整收敛因子的衰减曲线,提高 BGWO 算法的全局搜索能力,避免进入局部最优.

1 基础理论

1.1 F-score 算法

F-score 即 Fisher-score,是一种简单、快捷的特征选择模型. 在多分类的 F-score 特征选择模型[6]中,给定训练数据样本 $x^k \in R$,则训练样本 x^k 对应第 k 个特征的 F-score 定义为:

$$F_k = \frac{\sum_{i=1}^{c} \frac{n_i}{n} (m_i^k - m^k)^2}{\frac{1}{n} \sum_{i=1}^{c} \sum_{x \in \omega_i} (x^k - m_i^k)^2}, \tag{1}$$

其中,x^k 表示样本 x 在第 k 个特征上的取值,m_i^k 表示第 i 类样本在第 k 个特征上的取值的均值,m^k 表示所有类别的样本在第 k 个特征上的取值的均值. 需要说明的是,多分类的 F-score 通过类内方差和类间方差求得的 F_k 得分,可以很好反映特征对标签分类的贡献度. F_k 越高,对应特征的区别度也就越高.

1.2 Spearman 相关系数

Spearman 相关系数是衡量两个变量的依赖性的非参数指标[14]. 本文需要对肿瘤基因计算相互之间的 Spearman 相关系数,具体描述如下.

(1)计算原始肿瘤基因各样本之间特征的等级数,描述如下:

$$d_s = X_{is} - X_{js}, \tag{2}$$

其中,$s \in \{1,2,3,\cdots,n\}$,X_{is} 和 X_{js} 分别为第 s 个样本在第 i 个和第 j 个基因的等级排名,d_s 为第 s 个样本在第 i 个和第 j 个特征之间的等级差.

(2)计算两两基因之间的 Spearman 相关系数,描述如下:

$$sp(x_i;x_j)=1-6\sum_{s}^{n} d_s^2/n^2(n-1), \tag{3}$$

其中,x_i 和 x_j 表示第 i 个和第 j 个基因序列,n 表示基因 x_i 和 x_j 的样本个数.

Spearman 相关系数的值位于[-1,1]之间,$sp=1$ 表示两个特征呈现正相关,$sp=-1$ 表示两个特征呈现负相关,其绝对值越趋近于 1,两个特征之间的相关性越强.

1.3 二进制灰狼优化算法

二进制灰狼优化(BGWO)算法具有收敛性能较强、参数少、易实现等特点. 灰狼搜索猎物时会逐渐地接近猎物并包围它[10],其数学模型如下:

$$D=C\circ X_p(t)-X(t), \tag{4}$$

$$A=2a\circ r_1-a, \tag{5}$$

$$C=2r_2, \tag{6}$$

其中,t 为当前迭代次数,$\circ$表示 hadamard 乘积操作,A 和 C 是两个系数,X_p 表示猎物的位置,$X(t)$表示当前灰狼的位置,在整个迭代过程中收敛因子 a 由 2 线性降到 0,r_1 和 r_2 是[0,1]中的随机向量[10].

二进制灰狼优化过程的位置更新方式通过以下转换函数完成[4]:

$$V_p=\begin{cases}1, & \dfrac{1}{1+e^{-10(A_qD_p-0.5)}}>rand,\\ 0, & \text{其他},\end{cases} \tag{7}$$

其中,$p=\alpha,\beta,\delta,q=1,2,3,V_p$ 分别为 α 狼、β 狼和 δ 狼当前距离猎物位置的距离的二进制表示,D_p 和 A_q 可以通过式(4)和式(5)计算,$rand$ 表示[0,1]内的随机值.

$$x_q=\begin{cases}1, & (x_p+V_p)\geqslant 1,\\ 0, & \text{其他},\end{cases} \tag{8}$$

其中,$q=1,2,3,x_q$ 分别表示进行捕猎或者包围后 α 狼、β 狼和 δ 狼的位置. 本文采用的交叉策略为:

$$x_d=\begin{cases}a_d, & rand<\dfrac{1}{3},\\ b_d, & \dfrac{1}{3}\leqslant rand<\dfrac{2}{3},\\ c_d, & \text{其他},\end{cases} \tag{9}$$

$$X_i^{t+1}=\text{Crossover}(x_1,x_2,x_3), \tag{10}$$

其中,a_d、b_d 和 c_d 分别记录对应随机数范围的位置下标,从 3 个位置中选取对应数值完成交叉,实现本次循环的种群位置更换.

2 肿瘤基因选择方法

2.1 改进的 F-score

本文引入 Spearman 系数考虑特征之间的相似性,提出改进的 F-score(improved F-score method with the Spearman correlation coefficient,IFSSP),使用 IFSSP 算法对肿瘤基因数据集进行数据预处理,在不减少有效特征的前提下,去除无关基因降低维度. 具体来讲,为了解决没有充分考虑特征之间的相关性而造成多个相似特征得分均位居前列,其他特征排序靠后,影响特征选取的问题,根据 Spearman 相关系数绝对值描述冗余性,引入权重系数,构建特征与类别之间相关性和特征之间冗余度,通过冗余性惩罚待选择特征的相关性,改进 F-score 算法,有效反映特征在标签类别上的重要程度,进而提高初选特征的有效度和准确度.

在给定的数据集中,假设筛选的特征集为 S,则特征和类别之间的相关性表示所有 F-score 的平均值,描述如下:

$$F(S)=\max\left[\frac{1}{|s|}\sum_{x_i\in S}F_i\right], \tag{11}$$

其中,F_i 为特征集 S 中的特征 x_i 的 F-score 值,$|s|$为特征集 S 中的特征数目.

在给定的数据集中,假设筛选的特征集为 S,由于特征之间无论是接近正相关还是负相关,在相关性大的情况下,两个特征总会存在冗余,因此特征之间的冗余性表示为所有特征之间的 Spearman 相关系数的绝对值的平均值,描述如下:

$$SP(S)=\min\left[\frac{1}{|s|^2}\sum_{x_i,x_j\in S}spa(x_i;x_j)\right], \tag{12}$$

其中,$spa(x_i;x_j)$为 x_i 和 x_j 两个特征序列的 Spearman 相关系数的绝对值.

定义 1　为了缓解 $F(S)$和 $SP(S)$值相差较大,且占据主要地位,用于减少特征之间冗余度的 $SP(S)$不起作用情况的发生,权重系数定义如下:

$$\beta=\frac{\overline{spa}}{\overline{spa}+\bar{F}}, \tag{13}$$

其中,$\overline{spa}$为特征之间的 Spearman 相关系数的绝对值的平均值,$\bar{F}$ 为各个特征的 F-score 的平均值.

定义 2　对于已经筛选的特征集 S,在特征和类别相关性的基础上,当考虑了特征之间冗余性时,特征子集 S 的适应度函数定义如下:

$$FSP(S)=\max[\beta F(S)-(1-\beta)SP(S)]. \tag{14}$$

为了方便后续的筛选,采用增量方式表示,假设已经选择 $m-1$ 个特征的集合为 S_{m-1},则需要从余下的特征集合中,选取与类别相关性最大且冗余性最小的特征,描述如下:

$$FSP_p=\max_{x_i\in X-S_{m-1}}\left[\beta F_i-\frac{1-\beta}{m-1}\sum_{x_j\in S_{m-1}}spa(x_i;x_j)\right]. \tag{15}$$

孙林等[15]通过改进的 F-score 算法显著降低了多标记数据的特征维度,并提高了其分类效率. 吴迪等[16]通过引入类间散度度量和最大互信息改进 F-Score,并取得了显著的降维效果. 由以上分析可知,通过使用改进的 F-score 算法进行特征过滤,能够筛选若干排名靠前的特征作为预选特征子集,达到初步降维的目的.

2.2　改进的 BGWO 算法

为了加强 BGWO 算法的全局搜索能力和局部搜索能力,有效提高收敛速度和计算精确度,从以下两个方面进行优化:

(1)改进的收敛因子

在 BGWO 算法中,A 是主要影响到广度搜索和深度搜索的变量,当$|A|>1$ 时,灰狼群体将扩大包围圈,进行全局搜索;当$|A|<1$ 时,灰狼群体将缩小包围圈,完成对猎物的攻击,进行局部搜索[12]. 由式(5)可知,A 的变化是由收敛因子 a 决定的,且在整个迭代过程中收敛因子 a 由 2 线性降到 0,当 $a=1$ 时,算法全局搜索和局部搜索能力均等. 因此,可以通过改进 a 来加强全局搜索能力,避免该算法进入局部最优.

定义 3　改进的收敛因子 a 的衰减曲线描述如下:

$$a=2\mathrm{e}^{-4.5\left(\frac{iter}{Iter_{\max}}\right)^6}, \tag{16}$$

其中,$iter$ 为当前次的迭代次数,$Iter_{\max}$为最大迭代次数.

王梓辰等[17]通过设计非线性收敛因子来平衡鲸鱼优化算法的全局搜索和局部开发的能力,避免其陷入局部最优. 崔鸣等[18]设计了非线性收敛因子解决灰狼优化算法容易陷入局部最优的缺点. 基于以上分析,改进收敛因子以改变其更新方式,可以加强 BGWO 算法的全局搜索能力和局部开发能力.

(2)改进的狼群初始化

为了解决 BGWO 算法中初始化随机性广泛,进行迭代寻优时间开销大的问题,根据 FSP_p 值的大小计算各个特征的选择概率,按照概率进行特征选择,有效降低模型的时间开销.

定义 4　由式(15)计算出当前数据集各个特征的 FSP_p 值,在此基础上计算选择机率 PP_i,描述如下:

$$PP_i=\frac{FSP_i-FSP_{\min}}{FSP_{\max}-FSP_{\min}}, \tag{17}$$

其中,$FSP_{\max}$和 $FSP_{\min}$分别是当前特征中的最大值和最小值.

定义 5　计算各个特征被选择的概率 P_i,描述如下:

$$P_i = PP_i \Big/ \sum_j^n PP_j, \tag{18}$$

其中,PP_i 可以由式(17)计算得到.

(3)适应度函数

在特征选择中,适应度函数的优良直接影响分类结果的分类准确率,本文综合考虑分类准确率和筛选特征数目,确立新的适应度函数为:

$$fitness = \alpha(1-ACC_R) + \gamma \frac{|R|}{|C|}, \tag{19}$$

其中,$fitness$ 为当前的适应度值,$ACCR$ 为选择当前 R 个特征时的准确率,R 为筛选出来的特征数目,C 为特征总数,α 和 γ 为加权系数且 $\alpha+\gamma=1$.

2.3 算法描述

首先,计算每个特征的 F-score 和特征之间的 Spearman 相关系数的绝对值;然后,根据式(12)计算权重系数,由式(16)计算特征权重,依据重要性进行排序,筛选初选特征子集;最后,使用改进的 BGWO 算法进行特征选择,得到最优特征子集. 由此,设计一种基于改进 F-score、Spearman 相关系数和 BGWO 算法的肿瘤基因选择算法(tumor gene selection algorithm with the improved F-score, Spearman correlation coefficient, and BGWO, TGSFSB),其详细伪代码描述如下:

算法 1 TGSFSB 算法

输入:给定的基因数据集包含 m 个基因和 n 个样本,初选基因数目 p

输出:最优选特征子集 Pbest

/ * 初选特征的 IFSSP 算法 * /

Step 1. 根据式(1)计算所有特征的 $F_i(1 \leqslant i \leqslant m)$,由式(3)计算特征间的 Spearman 相关系数

Step 2. While $i<=p$, where $i=1,2,\cdots,p$

Step 3. 根据式(13)计算权重系数 β,由式(15)计算特征和标签之间的 FSP 值

Step 4. 根据特征和标签之间的 FSP 值,筛选出值最大的基因加入集合 S

Step 5. End While

Step 6. 得到初次筛选出的特征子集 S_p

/ * 使用改进的 BGWO 算法进行精选特征 * /

Step 7. 初始化种群数量 particle,种群维度 dim,最大迭代次数 Itermax,三头狼的得分 score_alpha、score_beta、score_delta,三头狼的位置 X_alpha, X_beta, X_delta;由式(17)随机初始化灰狼个体的位置 X

Step 8. While $iter<=Itermax$

Step 9. 根据式(19)计算当前适应度 score

Step 10. If $score<score\_alpha$ then 更新 α 狼的得分和位置

Else If $score>score\_alpha$ and $score<score\_beta$ then 更新 β 狼的得分和位置

Else If $score>score\_beta$ and $score<score\_delta$ then 更新 δ 狼的得分和位置

Step 11. End IF

Step 12. 根据式(16)计算收敛因子 a,随机化参数 $r1$、$r2$ 和 $r3$,根据式(4)-(6)分别计算 A、C 和 D,根据式(7)和式(8)计算 V 和三匹狼的最新位置,根据式(9)和式(10)进行交叉更新种群 i 的狼群位置,更新最优选择的分数及最优位置

Step 13. $iter=iter+1$

Step 14. End While

Step 15. 得到组成最终的特征子集 Pbest

算法 1 的时间复杂度分析如下:Step 1 计算各特征和标签之间的 F-score 的时间复杂度为 $O(m)$,对各个标记计算赋权值的时间复杂度为 $O(m\log m)$;Step 2 至 Step 6 使用增量方法,筛选出初次筛选的特征子集 S_p,时间复杂度为 $O(mp)$;Step 8 到 Step 14 是优化后的 BGW 算法搜索过程,其中 Step 9 到 Step 11 计算适应度,更新三匹狼的得分和位置,时间复杂度为 $O(p)$,Step 12 更新狼群位置的时间复杂度为 $O(p)$,所

以 Step 8 到 Step 14 的时间复杂度为 $O(2iter*p+iter)$. 由此 TGSFSB 算法总的时间复杂度为 $O(m+m\log m+mp+2iterp+iter)$.

3　实验

3.1　实验设置

为了检验 TGSFSB 算法的有效性和优越性,本文采用十则交叉验证方法[19],结合最终选取基因数和分类准确率[20]来评判各个对比算法的分类性能. 本文实验结果均选用各分类器下的十则交叉验证最优结果. 在 UCI 数据库中选取了 9 个公共的基因表达谱数据集进行实验,具体的数据集描述见表 1. 在对比算法表格中,粗体表示最优结果. 同时,本文选择被广泛使用的 SVM、KNN 和 RF 作为分类器,通过分类后的结果验证 TGSFSB 算法的实验效果. 仿真实验用到的计算机配置为 Windows 10、Intel Core i7-9750H、2.60 GHz、16 GB 内存,并采用 Pycharm 2022 和 jupyter lab 实现代码编写.

表 1　9 个基因表达谱数据集信息

Table 1　9 gene expression profile datasets information

序号	名称	基因数	样本数	样本类别	序号	名称	基因数	样本数	样本类别
1	Colon	2 000	62	2	6	GLI	9 430	85	2
2	Leukemia	7 129	72	2	7	Medulloblastoma	5 893	34	2
3	Lung	12 600	203	2	8	CNS	7 129	60	2
4	DLBCL	5 469	77	2	9	SRBCT	2 308	54	2
5	Brain	10 509	102	2					

3.2　改进 F-score 前后的实验结果分析

从表 1 中选择 8 个代表性数据集,在分类准确率指标下,使用 SVM、KNN 和 RF 作为分类器,对原始 F-score 和 IFSSP 算法的分类性能进行结果对比,具体分析如下:在 Colon 数据集上,在 SVM 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其当 N 取 40 时,IFSSP 算法的分类准确率比原始 F-score 算法提高了 4.999%;在 KNN 分类器下,IFSSP 算法的分类准确率与原始 F-score 算法近似相同;在 RF 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其是 N 取 50 时,IFSSP 算法比原始 F-score 算法提高了 1.901%. 在 Leukemia 数据集上,在 SVM 分类器下,IFSSP 算法的分类准确率均与原始 F-score 算法相等;在 KNN 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其当 N 取 90 时,IFSSP 算法比原始 F-score 算法提高了 2.679%;在 RF 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其是 N 取 150 时,IFSSP 算法比原始 F-score 算法提高了 2.678%. 在 Lung 数据集上,在 SVM 分类器下,IFSSP 算法与原始 F-score 算法效果均相同;在 KNN 分类器下,IFSSP 算法的分类准确率略优于原始 F-score 算法;在 RF 分类器下,IFSSP 算法与原始 F-score 算法效果近似一样. 这可能是由于 Lung 数据集中特征之间的冗余度较小造成的. 在 DLBCL 数据集上,在 SVM 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其当 N 取 50 时,提高了 4.999%;在 KNN 分类器下,IFSSP 算法的分类准确率略优于原始 F-score 算法,尤其当 N 取 50 时,提高了 6.607%;在 RF 分类器下,IFSSP 算法的分类准确率略优于原始 F-score 算法,尤其 N 取 30 时,IFSSP 算法比原始 F-score 算法提高了 6.429%. 在 Brain 数据集上,在 SVM 分类器下,IFSSP 算法的分类准确率略优于原始 F-score 算法,尤其当 N 取 100 时,IFSSP 算法比原始 F-score 算法提高了 1.909%;在 KNN 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其当 N 取 20 时,IFSSP 算法比原始 F-score 算法提高了 2.818%;在 RF 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其是 N 取 70 和 130 时,IFSSP 算法比原始 F-score 算法提高了 2.000%. 在 GLI 数据集上,在 SVM 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其当 N 取 110 时,分类准确率比原始 F-score 算法提高了 7.222%;在 KNN 分类器下,IFSSP 算法的分类准确率略均优于原始 F-score 算法;在 RF 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其是 N 取 30 时,IFSSP 算法比原始 F-score 算法提高了 2.222%. 在 Medulloblastoma 数据集上,在 SVM 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,且最后趋于平稳的定值,尤其当 N 取 100 时,IFSSP 算法比原始 F-score 算法提高了 5.833%;在 KNN 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法;在 RF 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其是 N 取 70 时,IFSSP 算法比原始 F-score 算法提高

了 3.333%. 在 CNS 数据集上,在 SVM 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其当 N 取 40 时,IFSSP 算法比原始 F-score 算法提高了 11.667%;在 KNN 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其当 N 取 150 时,IFSSP 算法比原始 F-score 算法提高了 3.333%;在 RF 分类器下,IFSSP 算法的分类准确率均优于原始 F-score 算法,尤其是 N 取 20 和 30 时,IFSSP 算法比原始 F-score 算法分别提高了 5.000%和 6.667%.

为了更深入证实 IFSSP 算法的有效性,使用 SVM、KNN 和 RF 作为分类器,以 3 个分类器的十则交叉验证平均最高得分作为其分类准确率. 将原始数据集的结果、采用原始 F-score 算法初选特征子集和采用 IFSSP 算法初选特征子集的分类准确率进行对比,如表 2 所示,分类结果对比证实了改进的 IFSSP 算法的有效性.

表 2　原始数据、F-score 算法和 IFSSP 算法的分类结果对比

Table 2　Comparison of classification results between raw data, F-score algorithm, and IFSSP algorithm

数据集	原始数据		原始 F-score 算法		IFSSP 算法	
	基因数	分类准确率	基因数	分类准确率	基因数	分类准确率
Colon	2 000	0.804 7	**90**	0.869 0	**90**	**0.885 7**
Leukemia	7 129	0.957 1	**90**	0.973 2	**90**	**0.985 7**
Lung	12 600	0.980 2	**70**	0.989 9	**70**	**0.994 9**
DLBCL	5 469	0.952 5	**150**	0.937 5	**150**	**0.960 7**
Brain	10 509	0.910 9	**100**	0.930 0	**100**	**0.930 9**
GLI	9 430	0.893 0	**120**	0.891 7	**120**	**0.927 8**
Medulloblastoma	5 893	0.750 0	**120**	0.808 3	**120**	**0.866 7**
CNS	7 129	0.600 0	**70**	0.800 0	**70**	**0.833 3**
SRBCT	2 308	0.930 0	**60**	**1.000 0**	**60**	**1.000 0**
Avg.	6 940.78	0.864 3	**96.67**	0.911 1	**96.67**	**0.931 7**

综上所述,IFSSP 算法在进行特征选择时,可以实现同时考虑相关性和冗余度,过滤掉大量无关基因和冗余基因,较传统的 F-score 算法分类准确率有所提高,展示出优异的分类性能.

3.3　不同算法的实验对比

本节首先选取了表 1 中的 9 个公共基因数据集,在 SVM、KNN、RF 分类器上,将 TGSFSB 算法与 IFSSP 算法、基于原始 BGWO 的特征选择算法(BGWO)进行选择的基因数和分类准确率的对比. 表 3 为 3 种分类器下 3 种不同算法在 9 个数据集上的分类结果对比.

从表 3 可以看出,TGSFSB 算法在这三种分类器下分别在 4 个、4 个和 1 个数据集上的分类准确率达到 100%,分别在 7 个、6 个、5 个数据集上高于 97.00%,在 9 个数据集上的平均分类准确率均高于 95.00%. 在 SVM 分类器下,TGSFSB 算法的平均分类准确率较 IFSSP 算法和 BGWO 算法分别提高了 5.89%和 5.26%;在 Medulloblastoma 数据集上,TGSFSB 算法的分类准确率较 IFSSP 算法和 BGWO 算法分别提高 10.83%和 10.00%,筛选特征数较 IFSSP 算法和 BGWO 算法分别减少了 110 个和 3 251 个;在 CNS 数据集上,TGSFSB 算法筛选特征数较 IFSSP 和 BGWO 分别减少了 60 个和 4 377 个,分类准确率较 IFSSP 和 BGWO 分别提升 16.67%和 23.33%. 在 KNN 分类器下,TGSFSB 算法的平均分类准确率较 IFSSP 算法和 BGWO 算法分别提高了 6.10%和 5.84%;在 Colon 数据集上,TGSFSB 筛选的特征数较 IFSSP 和 BGWO 分别减少了 74 个和 1 042 个,分类准确率较 IFSSP 和 BGWO 分别提升 9.75%和 6.43%;在 Brain 数据集上,TGSFSB 筛选的特征数较 IFSSP 和 BGWO 分别减少了 113 个和 6 388 个,分类准确率较 IFSSP 和 BGWO 分别提升 10.82%和 10.82%;在 Medulloblastoma 数据集上,TGSFSB 筛选的特征数较 IFSSP 和 BGWO 分别减少了 114 个和 3 506 个,分类准确率较 IFSSP 和 BGWO 分别提升 8.33%和 9.17%;在 CNS 数据集上,TGSFSB 筛选的特征数较 IFSSP 和 BGWO 分别减少 106 个和 4 277 个,分类准确率较 IFSSP 和 BGWO 分别提升 10.00%和 16.66%. 在 RF 分类器下,TGSFSB 算法的平均分类准确率较 IFSSP 算法和 BGWO 算法分别提高了 2.51%和 5.05%;在 Medulloblastoma 数据集上,TGSFSB 筛选的特征数较 IFSSP 和 BGWO 分别减少了 95 个和 3 347 个,分类准确率较 IFSSP 和 BGWO 分别提升 8.34%和 9.17%,效果最优. 上述实验结果证明 TGSFSB 算法具有良好的分类效果.

接下来,选择 4 个代表性数据集,在上述 3 种分类器上,TGSFSB 算法与其他 4 种算法(S-BPSO[21]、

IS-SGA[22]、TSLR[23]和 LS-CNN[24])进行实验指标的对比,表 4 展示了 5 种不同算法在 4 个数据集上的对比结果,其中,"-"表示所对比论文没有当前数据结果.

表 3　3 种分类器下 3 种不同算法在 9 个数据集上的分类结果对比

Table 3　Comparison of classification results of 3 different algorithms under 3 classifiers on 9 datasets

分类器	数据集	IFSSP		BGWO		TGSFSB	
		基因数	分类准确率	基因数	分类准确率	基因数	分类准确率
SVM	Colon	90	0.885 7	1 235	0.916 7	**16**	**0.935 7**
	Leukemia	90	0.987 5	3 447	0.985 7	**8**	**1.000 0**
	Lung	100	0.990 0	5 442	0.985 0	**3**	**1.000 0**
	DLBCL	150	0.960 7	2 786	**1.000 0**	**8**	**1.000 0**
	Brain	100	0.930 9	6 446	0.960 0	**8**	**0.990 0**
	GLI	100	0.881 9	4 980	0.904 2	**16**	**0.965 3**
	Medulloblastoma	120	0.866 7	3 351	0.875 0	**10**	**0.975 0**
	CNS	80	0.816 6	4 397	0.750 0	**20**	**0.983 3**
	SRBCT	60	**1.000 0**	1 082	**1.000 0**	**3**	**1.000 0**
	Avg.	98.89	0.924 4	3 685.11	0.930 7	**10.22**	**0.983 3**
KNN	Colon	80	0.852 5	1 048	0.885 7	**6**	**0.950 0**
	Leukemia	90	0.971 4	4 077	0.985 7	**8**	**1.000 0**
	Lung	70	0.994 9	6 188	0.995 0	**3**	**1.000 0**
	DLBCL	150	0.933 9	2 924	0.960 7	**12**	**1.000 0**
	Brain	120	0.871 8	6 395	0.871 8	**7**	**0.980 0**
	GLI	190	0.893 0	5 529	0.916 7	**8**	**0.952 8**
	Medulloblastoma	120	0.891 7	3 512	0.883 3	**6**	**0.975 0**
	CNS	120	0.833 3	4 291	0.766 7	**14**	**0.933 3**
	SRBCT	60	**1.000 0**	**1 185**	**1.000 0**	**3**	**1.000 0**
	Avg.	111.11	0.915 8	3 905.44	0.918 4	**7.44**	**0.976 8**
RF	Colon	80	0.852 5	1 048	0.885 7	**6**	**0.950 0**
	Leukemia	90	0.971 4	4 077	0.985 7	**8**	**1.000 0**
	Lung	70	0.994 9	6 188	0.995 0	**3**	**1.000 0**
	DLBCL	150	0.933 9	2 924	0.960 7	**12**	**1.000 0**
	Brain	120	0.871 8	6 395	0.871 8	**7**	**0.980 0**
	GLI	190	0.893 0	5 529	0.916 7	**8**	**0.952 8**
	Medulloblastoma	120	0.891 7	3 512	0.883 3	**6**	**0.975 0**
	CNS	120	0.833 3	4 291	0.766 7	**14**	**0.933 3**
	SRBCT	60	**1.000 0**	1 185	**1.000 0**	**3**	**1.000 0**
	Avg.	111.11	0.915 8	3 905.44	0.918 4	**7.44**	**0.976 8**

表 4　5 种不同算法在 4 个数据集上的对比结果

Table 4　Comparison results of 5 different algorithms on 4 datasets

算法	指标	DLBCL	Colon	Leukemia	Lung	Avg.	Std.
TGSFSB	分类准确率	**1.000**	0.9500	**1.000**	**1.000**	0.9875	0.02
	基因数	**8.0**	6.0	8.0	**3.0**	**6.25**	2.04
S-BPSO	分类准确率	**1.000**	0.908	0.973	0.972	0.963	0.03
	基因数	21.5	27.7	22.2	61.7	33.3	16.59
IS-SGA	分类准确率	0.948	0.855	0.971	**1.000**	0.944	0.05
	基因数	110.0	60.0	**3.0**	9.0	45.5	43.32
TSLR	分类准确率	—	0.938	0.937	0.964	0.946	**0.01**
	基因数	—	**5.0**	7.0	9.0	7.0	**1.63**
LS-CNN	分类准确率	0.980	**1.000**	0.990	—	**0.990**	**0.01**
	基因数	1 500.0	1 000.0	1 700.0	—	1 400.0	294.39

由表 4 的对比结果可知,在 DLBCL 数据集上,TGSFSB 算法和 S-BPSO 算法的分类准确率最高,均达到了 100%,选择的基因数较 S-BPSO 少 13.5 个,而 TGSFSB 算法的分类效果和筛选性能最佳;在 Colon 数据集上,LS-CNN 算法的分类准确率达 100%,但选择的基因数为 1000,筛选性能差,TGSFSB 算法的分类准确率位居第二,筛选出的基因数比分类准确率最高的 LS-CNN 算法少 994 个,有效降低基因维度,比最少

的 TSLR 算法多 1 个,但是分类准确率提高了 1.20%;在 Leukemia 数据集上,TGSFSB 算法的分类准确率最高,达到了 100%,选择的基因数较最少的 IS-SGA 算法多了 5.0 个,但是分类准确率提高了 6.30%;在 Lung 数据集上,TGSFSB 算法和 IS-SGA 算法的分类准确率最高,均达到 100%,但选择的基因数分别较 IS-SGA 减少 6.0 个. 综合来看,TGSFSB 算法的分类效果和筛选性能最佳. 观察表 4 的 Avg.可得,TGSFSB 算法平均选择的基因数为 6.25,远低于另外 4 种算法,同时平均分类准确率比最优的 LS-CNN 算法低了 0.05%,但是在平均选择的基因数上减少了 1 393.75 个;观察 Std.可得,TGSFSB、TSLR、LS-CNN 这 3 种算法的分类准确率波动均小于或等于 0.02,波动较小,但 TGSFSB 算法的平均分类准确率优于其余两者. 综上所述,TGSFSB 算法能够获取最优基因子集,同时也展现出良好的的分类性能.

4 结论

本文利用 F-score 和特征之间的相关性,综合 Spearman 相关系数改进 F-score 算法,并在此基础上优化 BGWO 算法收敛因子的衰减曲线和初始化方法,提出了一种基于改进的 F-score 和 BGWO 的肿瘤基因选择方法. 首先,计算每个特征的 F-score 值和特征之间的 Spearman 相关系数的绝对值;然后,计算权重系数,从而得出各个特征的权重值,依据重要性进行排序,选出初选特征子集;最后,使用改进的 BGWO 算法对基因特征进行选择,得到终选特征子集. 在 9 个肿瘤基因数据集上测试所设计算法,进行分类准确率和选择基因数的对比,结果显示其有效提升了分类效果. 但是,在与其他算法对比时,本文算法的表现并不是全部达到最优. 因此,在未来的研究工作中,将结合机器学习最新理论,进一步优化肿瘤基因选择算法.

[参考文献]

[1] 吴辰文,纪海斌. 混合 mRMR 和改进磷虾群的肿瘤基因特征选择算法[J]. 西北大学学报(自然科学版),2022,52(2):262-269.

[2] 孙林,徐枫,李硕,等. 基于 ReliefF 和最大相关最小冗余的多标记特征选择[J]. 河南师范大学学报(自然科学版),2023,51(6):22-30.

[3] 马超. 基于 FCBF 特征选择和集成优化学习的基因表达数据分类算法[J]. 计算机应用研究,2019,36(10):2986-2991.

[4] 王琛,董永权. 基于二进制灰狼优化的特征选择及文本聚类[J]. 计算机工程与设计,2021,42(9):2526-2535.

[5] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine learning,2002,46:389-422.

[6] 谢娟英,王春霞,蒋帅,等. 基于改进的 F-score 与支持向量机的特征选择方法[J]. 计算机应用,2010,30(4):993-996.

[7] 谢娟英,郑清泉,吉新媛. F-score 结合核极限学习机的集成特征选择算法[J]. 陕西师范大学学报(自然科学版),2020,48(2):1-8.

[8] 吴晓燕,刘笃晋. 基于樽海鞘群与粒子群混合优化算法的特征选择[J]. 重庆邮电大学学报(自然科学版),2021,33(5):844-850.

[9] 秦喜文,王芮,于爱军,等. 基于 F-score 的特征选择算法在多分类问题中的应用[J]. 长春工业大学学报,2021,42(2):128-134.

[10] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer[J]. Advances in engineering software,2014,69:46-61.

[11] EMARY E,ZAWBA H M,HASSANIEN A E. Binary grey wolf optimization approaches for feature selection[J]. Neurocomputing,2016,172(8):371-381.

[12] 陈长倩,慕晓冬,牛犇,等. 结合高斯分布的改进二进制灰狼优化算法[J]. 计算机工程与应用,2019,55(13):145-150.

[13] 邢燕祯,王东辉. 一种基于收敛因子改进的灰狼优化算法[J]. 网络新媒体技术,2020,9(3):28-34.

[14] 王伟,吕婷婷,周晓冰. 河南 5A 级景区网络关注度时空演变特征与影响因素[J]. 河南师范大学学报(自然科学版),2023,51(2):70-78.

[15] 孙林,马天娇,薛占熬. 基于 Fisher score 与模糊邻域熵的多标记特征选择算法[J/OL]. 计算机应用:1-12[2023-08-18]. https://kns-cnki-net.webvpn.las.ac.cn/kcms/detail/51.1307.tp.20230214.1544.002.html.

[16] 吴迪,郭嗣琮. 改进的 Fisher Score 特征选择方法及其应用[J]. 辽宁工程技术大学学报(自然科学版),2019,38(5):472-479.

[17] 王梓辰,窦震海,董军,等. 多策略改进的自适应动态鲸鱼优化算法[J]. 计算机工程与设计,2022,43(9):2638-2645.

[18] 崔鸣,靳其兵. 基于 Levy 飞行策略的灰狼优化算法[J]. 计算机与数字工程,2022,50(5):948-952,958.
[19] 汪丽丽,邓丽,余玥,等. 基于 Spark 的肿瘤基因混合特征选择方法[J]. 计算机工程,2018,44(11):1-6.
[20] SUN L,WANG L Y,DING W P,et al. Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets[J]. IEEE transactions on fuzzy systems,2021,29(1):19-33.
[21] YANG J,LIU Y L,FENG C S,et al. Applying the Fisher score to identify Alzheimer's disease-related genes[J]. Genetics & molecular research gmr,2016,15(2):19-28.
[22] SALEM H,ATTIYA G,EL-FISHAWY N. Classification of human cancer diseases by gene expression profiles[J]. Applied soft computing,2016,50:124-134.
[23] ALGAMAL Z Y,LEE M H.A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification[J]. Advances in data analysis and classification,2019,13(3):753-771.
[24] SHAH S H,IQBAL M J,AHMAD I,et al. Optimized gene selection and classification of cancer from microarray gene expression data using deep learning[J]. Neural computing and applications,2020,(3/4):1-12.

[责任编辑:杜忆忱]