基于混合进化算法的特征选择方法研究

高慧敏[1],王云鹤[2],卞 闯[1],李向涛[1]
(1. 吉林大学人工智能学院,吉林长春 130000;2. 河北工业大学人工智能与数据科学学院,天津 300401)

摘 要: 特征选择(Feature Selection,FS)是一种有效的数据预处理方法,它可以通过选择高维数据中一组具有高相关性和低冗余性的特征,从而解决数据冗余引起的维数灾难.目前许多计算方法已经被应用于求解FS问题,其中基于教与学优化(Teaching and Learning-based Optimization Algorithm,TLBO)的特征选择模型由于其高效的全局搜索能力受到越来越多学者的关注.然而,随着数据规模的不断扩大,这些算法所具有的模型不稳定、模型精确度低和局部搜索能力差等局限性,使算法的研究逐步陷入困境.为解决上述问题,本文提出了融合教与学优化算法与局部搜索方法(Local Search,LS)的混合进化Wrapper算法模型(Teaching and Learning-based Optimization- Local Search Algorithm,TLBOLS).首先,由于传统的教与学优化算法不能直接用于求解特征选择问题,算法在初始化阶段将实数型编码转为二进制编码,然后为保证种群的多样性,在教阶段引入最差个体重启机制,并针对进化班级过程中学习者与教学者两种身份采用不同值的TF值,提出二进制的教与学特征选择算法(Binary Teaching and Learning-based Optimization-Local Search Algorithm,BTLBOLS).随后,提出结合多操作的局部搜索方法和变邻域搜索逐渐增强扰动力度,提高整个种群的个体质量.为优化特征选择结果,BTLBOLS利用综合评价指标作为目标函数指导整体进化过程.实验选取45个高维癌症基因表达数据集进行测试并与十种特征选择算法相比,实验结果表明,相比其他算法,BTLBOLS在分类准确率和特征个数上都具有一定优势,算法分类性能有效提高.

关键词: 教与学优化算法;局部搜索;新型Wrapper混合特征选择算法;特征选择;分类;基因表达数据
基金项目: 国家自然科学基金(No.62076109)
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112(2023)06-1619-18
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210399

Research on Feature Selection Based on Hybrid Evolutionary Algorithm

GAO Hui-min[1], WANG Yun-he[2], BIAN Chuang[1], LI Xiang-tao[1]
(1. *School of Artificial Intelligence*, *Jilin University*, *Changchun*, *Jilin* 130000, *China*;
2. *School of Artificial Intelligence*, *Hebei University of Technology*, *Tianjin* 300401, *China*)

Abstract: Feature selection (FS) is an effective data pre-processing method that solves the dimensionality disaster caused by data redundancy by selecting a set of features with high relevance and low redundancy in high-dimensional data. Many computational methods have been applied to solve the FS problem, among which the teaching and learning-based optimization algorithm (TLBO) feature selection model has received increasing attention from scholars due to its efficient global search capability. However, with the increasing size of data, the limitations of these algorithms, such as model instability, low model accuracy and poor local search ability, have gradually put the research of the algorithms into difficulties. To address these problems, this paper proposes a hybrid evolutionary Wrapper algorithm model (Teaching and Learning-Based Optimization- Local Search algorithm,TLBOLS) that integrates teaching-learning optimization algorithms with local search methods. Firstly, the algorithm converts the real-type coding to binary coding in the initialization phase, then introduces the worst individual restart mechanism in the teaching phase, and proposes a binary teaching-learning feature selection algorithm for the evolutionary class process using different values of TF values for the two identities of learners and pedagogues (Binary Teaching and Learning-Based Optimization- Local Search algorithm, BTLBOLS). Subsequently, a local search method combining multiple operations and variable neighborhood search is proposed to gradually enhance the perturbation strength and improve the individual quality of the whole population. To optimize the feature selection results, BTLBOLS utilizes a comprehensive evaluation metric as an objective function to guide the overall evolutionary process.

收稿日期:2021-03-27;修回日期:2022-04-29;责任编辑:王天慧

Forty-five high-dimensional cancer gene expression datasets are selected for testing and compared with ten feature selection algorithms, and the experimental results show that compared to other algorithms, the BTLBOLS has certain advantages in terms of classification accuracy and number of features, which effectively improves the algorithm classification performance.

Key words: teaching and learning-based optimization algorithm; local search; new Wrapper hybrid feature selection algorithm; feature selection; classification; gene expression data

Foundation Item(s): National Natural Science Foundation of China (No.62076109)

1 引言

特征选择是一种有效去除数据集中冗余、低相关特征的数据预处理方法,有望解决数据冗余引起的维度灾难,并提高后续学习如分类、聚类等数据挖掘任务的精度[1]. 具体过程如下:设高维数据具有N维特征即$\boldsymbol{p}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N)$,经过一系列特征评价操作后得到新数据$\boldsymbol{p}_{\text{new}}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_M)$,其中$M<N$,该过程称为特征选择. 特征选择之所以是一个难解决问题,主要因为搜索空间会随着数据集特征的个数增长呈指数级增长,例如,若数据具有N个特征,那么其非空子集就有2^N-1个. 由此可见,随着数据规模的不断增大,穷举法会显得无能为力. 因此,开发高效的特征选择方法已成为高维数据挖掘中必不可少的一环. 常见的特征选择方法可分为三类:过滤式(Filter)[2]、包裹式(Wrapper)[2]、嵌入式(Embedded)[3].

(1) 过滤式(Filter)方法仅依赖于特征本身特性选择,不涉及任何分类器或分类算法. 根据是否考虑特征之间的相互作用可以分为单变量Filter和多变量Filter. 单变量Filter通常采用某种度量方式测量特征与标签之间的关系,例如T-检验[4]和reliefF[5]方法等;多变量考虑特征之间的相互作用和数据集的冗余度,常见方法有MRMR(Minimum Redundancy-Maximum Relevance)[6,7]和CSF(Correlation-based Feature Selection)[8]等.

(2) 包裹式(Wrapper)方法将分类算法或分类器的精度作为特征子集评价标准,在特征子集构成的搜索空间中搜索可以使所用分类器精度最高的特征子集. 该方法包含两个要素:搜索策略和分类器. 两者相辅相成:分类器指导搜索策略,搜索策略产生的特征子集进入分类器训练. Wrapper模型可概括为

$$\mathbf{Sub}_{\text{best}}=\text{argmax}\{\text{Per}(\boldsymbol{D}(\boldsymbol{S}),C)\}.$$

其中,$\mathbf{Sub}_{\text{best}}$为一定次数迭代以后的最优特征子集;$\boldsymbol{D}(\boldsymbol{S})$为所有特征子集;$C$为搜索过程中所使用的分类器,Per为分类器的精度.

(3) 嵌入式(Embedded)则将分类器与特征选择融为一体,在训练分类器的同时进行特征选择[9],一般分为两种方式:增加或减少特征影响分类器目标函数进行特征选择;将特征个数信息作为正则项加入分类器目标函数进行选择的同时保证特征个数最少. 其中Guyon等人[10]在2002年提出的SVMRFE最具代表性.

三种方法的比较:从运行时间看,Filter方法所需时间最短;在Embedded方法中,以SVMRFE为例,特征选择时,SVMRFE依次移除对SVM目标函数影响最小的特征,由于移除第i个特征之前的目标函数与移除它之后的目标函数差值为w_i^2,故SVMRFE仅需要训练一次SVM即可以判断出移除任一特征对SVM目标函数的影响;而在Wrapper方法中,每移除一个特征就需要将特征子集重新输入分类器中进行分类学习判断其影响,故Wrapper要比Embedded的运行时间长.

同时,根据在特征选择过程中搜索策略的不同,还可将特征选择分成启发式算法和进化计算两类. 启发式算法尝试缩小搜索空间,更关注于可能产生最优解的特征组合. SFS(Sequential Forward Selection)[11]和SBS(Sequential Backward Selection)[12]是两种经典的爬山法,但这类算法易陷入局部最优解. 为克服该缺点,SFFS和SBFS算法被相继提出[13]. 近年来,研究学者提出采用支持向量机的启发式搜索策略进行特征选择. 它本身是一种有监督学习,主要包括RFR(Recursive Feature Replacement)和RFE(Recursive Feature Elimination). 例如RFE方法基于SFS,以特征集的交叉验证错误率作为评价指标;SVM-RFE方法[14]在特征排序过程中使用SVM判别函数信息.

进化优化的出现,给问题的解决带来新的希望. 进化算法是模仿生物自然选择与自然进化的一种新兴智能优化系统,相较于传统方法,进化算法可以进行个体间信息交换. 目前多种基于进化计算的特征选择方法已提出如文献[15]利用进化优化以及大规模种群NSGA-II范式解决同时存在区间显式指标和模糊隐式指标的高维混合指标优化问题;文献[16]利用均匀设计来构造新的高效进化算法,攻克有效地搜索解空间,保持种群的多样性,减小计算量的难点;文献[17]将进化计算应用到隐马尔柯夫模型(Hidden Markov Model, HMM)的训练中,提出将传统算法和进化计算相结合的混合算法既保证了全局搜索又实现了快速收敛;Jimenez等人提出利用ENORA (Evolutionary Non-dominated Radial slots based Algorithm)进行多目标进化特征选择[18],并将多目标进化算法与基于模糊规则的

学习相结合,进行高维数据特征选择[19];Mafarja 等人[20]将 GOA(Grasshopper Optimization Algorithm)与选择算子和进化种群动力学结合进行特征选择;Taradeh 等人[21]使用GSA(Gravitational Search Algorithm)进行特征选择;BaharehNakisa[22]提出利用进化计算对高维 EEG(Electro Encephalon Gram)信号进行特征选择;Ghosh 等人[23]提出根据问题自适应参数的 DE(Differential Evolutionary Algorithm)解决特征选择问题. 同时,研究人员对许多混合进化算法进行研究. 例如,Khushab 等人[24]将 DE 用于根据 ACO 获得的解决方案搜索最佳特征子集,所提组合增强了搜索过程的探索和开发能力;Zorarpacı 等人[25]利用 ABC(Artificial Bee Colony Algorithm)为 DE 提供一种新的二进制突变机制进行特征选择;Allaoui 等人[26]将 CSA(Crow Search Algorithm)混合局部搜索算法,对 DNA(Deoxyribonucleic Acid)片段进行选择;Shukla 等人[27]融合 TLBO(Teaching and Learning-Based Optimization Algorithm)和 GSA(Gravitational Search Algorithm),在教学阶段纳入引力搜索机制并与 mRMR(minimum Redundancy Maximum Relevance)结合,对高维癌症基因数据集进行特征选择;文献[28]以达尔文自然进化论为灵感的神经进化方法训练以种群为基础的深度学习模型,通过突变、重组等操作进化,实现自动地、逐步地构建神经网络并最终选择出性能最优的深度学习模型.

2011 年,Rao 等人[29]研究了一种新的元启发式算法,称为 TLBO(Teaching and Learning-Based Optimization Algorithm). 该算法模仿一个包含若干行为的班级,学生可以在老师和其他同伴中获取知识,在初始时将最佳学生设置为班级老师. 由于其较强的全局搜索能力,该方法已被广泛应用于不同的领域. 2016 年,Wang 等人[30]在 TLBO 算法中引入经验信息 EI 和差异突变,在进化过程中保证种群的多样性并提高算法的收敛速度和准确性,且对 46 种基准函数进行评估,结果表明 EI-TLBO 可以达到较高性能. 2011 年,Rao 等人[29]将 TLBO 算法应用于机械设计问题的优化并在具有不同特性的五个不同约束基准测试功能、四个不同基准机械设计问题和六个具有实际应用的机械设计优化问题上进行了测试,证明 TLBO 方法的有效性. 2012 年,Rao 等人[31]就非线性优化问题,引入 TLBO 算法进行优化,在许多具有不同特征的基准问题上测试了该方法的有效性,并将结果与其他进化算法进行比较,验证了 TLBO 算法在解决非线性问题的有效性. Zou 等人[32]将 TLBO 算法与动态组策略 DGS 结合,维持种群多样性并避免过早收敛,在 10、30 和 50 维上对 18 个基准函数进行实验,证实 DGSTLBO 在解决全局优化问题时的有效性. Ghasemi 等人[33]将改进后的教与学算法 MTLA 与双差分进化算法 DDE 结合,解决电力系统中最优无功功率分配问题,并将算法应用于性能评估和验证目的的 IEEE 14 总线、IEEE 30 总线和 IEEE 118 总线电源系统上的 ORPD 问题,提供更好的解决方案. González-Álvarez 等人[34]将多目标教与学算法 MO-TLBO 应用于生物信息学领域,试图解决主题发现问题,并通过一组十二个不同生物学实例证实了该方法与其他多目标进化算法相比时具有更优性能. 以上均充分说明了 TLBO 的高效性.

为解决上述提出的问题,本文基于 TLBO 算法提出一种新的混合进化算法,克服传统算法的局限性,选择高表达癌症亚型基因子集. 这是 TLBO 和局部搜索的首次结合,形成了混合 Wrapper 模型,该模型解决了原算法的局限性. 此外,本文采用二进制方式对种群个体进行编码,利用综合适应度函数指导进化,同时解决基因选择问题和适应性评估问题. 在实验部分,采用 45 个癌症数据集与多种进化特征选择算法相比较. 此外,它还具有全局搜索和处理大量数据的能力,可提供合理的解决方案. 本文的贡献总结如下:

(1)提出将连续空间的教与学算法应用于特征选择问题,对庞大的癌症基因数据集进行特征识别,作为诊断依据并提高分类器的分类精度.

(2)为离散化特征选择问题,引入二进制编码机制;且为避免算法陷入局部最优解,引入新的班级更新机制,如最差个体重启策略;教阶段根据教师和学生不同情况采用不同 TF 值更新班级.

(3)为进一步提高可行解的质量,将 BTLBO 与局部搜索方法结合形成新型 Wrapper 特征选择算法,不仅提高了 BTLBO 算法的收敛性,还平衡了局部搜索和全局搜索.

(4)为验证算法的有效性,本文采用 35 个基因表达数据集. 实验结果表明,本文提出的算法可以选择出具有高度判别力的基因子集,且与其他特征选择算法相比,所提算法可得较优分类精度.

2　Teaching and Learning-Based 算法概述

教与学优化算法(Teaching and Learning-Based Optimization Algorithm,TLBO)是由 Rao 等人在 2011 年提出的一种基于群体的启发式优化算法[29],它最初被用于解决连续空间中的非线性问题. TLBO 主要基于师生互动效果,每个个体都是一个可行的解决方案. 该算法模仿一个包含若干行为的班级,学生可以向老师和其他同伴学习知识,而开始时将最佳学生设置成班级老师:其由于较强的全局搜索能力,通常用于解决非线性

问题. 此外,该算法在收敛速度、运行时间及内存占用空间等方面都具有出色的性能. 本文基于传统的教与学优化算法提出二进制TLBO(BTLBO)算法,以解决离散的特征选择问题.

在BTLBO算法中,每个个体都代表问题的一个可行解,即一个特征子集. 每个可行解都是一个长度为特征数的二进制向量,向量中每个元素都代表一个特征,0表示对应特征被摒弃,1表示相应特征被选择到特征子集里参与后续学习任务,具体如图1所示. 随后,将这些可行解所表示的特征子集放入分类器中进行训练,从而获得准确率. BTLBO算法的具体流程见图2,观察可得,该算法一共分为三个阶段:初始化班级、提高班级平均水平的教学阶段及加强个体水平的学习阶段. 该算法以初始班级为起点,经过教阶段与学阶段两个进化步骤,不断更新班级,找到最优特征子集.

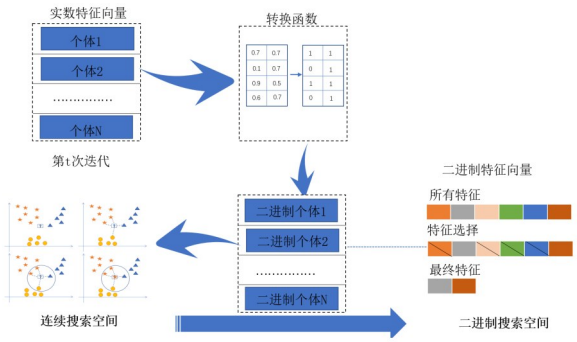

图1 特征的二进制向量表示

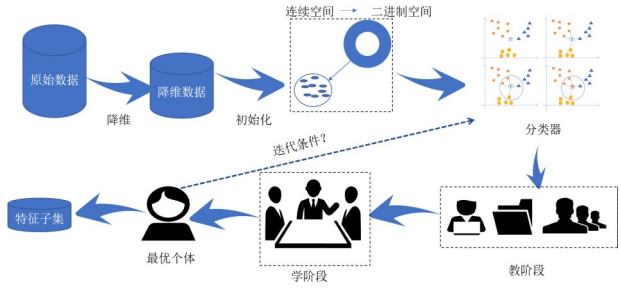

图2 BTLBO算法流程图

初始化阶段包含N个个体的班级,每个个体都是一个D维二进制向量,即一个可行解. 所有可行解与非可行解共同构成一个搜索空间. 初始化过程可描述为

$$\mathrm{pop}_i = L_k + r \times \left(U_k - L_k\right) \tag{1}$$

其中,pop_i是班级中第i个个体,L_k和U_k分别是变量的上界和下界,r是一个随机分布在[0,1]的实数,k代表第k维. 为使算法可直接求解特征选择问题,随机生成实数型特征向量后,将每个维度的数值与阈值δ进行比较,得到新的二进制向量,将每个个体代表的特征子集放入分类器进行训练,得到每个个体的Acc和fitness,至此初始化阶段全部完成.

在教阶段根据初始化阶段得到所有个体分类性能,选出最优个体作为教师,为保证种群优良性,将最差个体按照式(1)进行重启. 随后,教师向学生输入知识,学生均以一定概率向教师方向移动,更新整个班级并以此来提高班级平均水平. 具体可描述为

$$\mathrm{pop}_i^{t+1} = \mathrm{pop}_i^t + r \times \left(\mathrm{tea}_i^t - \mathrm{TF} \times \mathrm{Mean}_i^t\right) \tag{2}$$

其中,tea_i^t表示第t次迭代中的教师;Mean_i^t为第t次迭代中整个班级的平均值;TF为教师参数,是[1,2]中的随机数. 需要注意的是,TF并非算法里的参数,只起到对班级平均值轻微干扰的作用. 在本文中,针对班级更新时教师和学生两种情况,采用不同TF值,即

$$\mathrm{TF} = \begin{cases} 1, & \mathrm{pop}_i = \mathrm{pop}_{\mathrm{tea}} \\ \dfrac{f\left(\mathrm{pop}_i\right)}{f\left(\mathrm{pop}_{\mathrm{tea}}\right) - f\left(\mathrm{pop}_i\right)}, & \mathrm{pop}_i \neq \mathrm{pop}_{\mathrm{tea}} \end{cases} \tag{3}$$

其中,f为特征选择过程中目标函数. 整个班级被更新后,将新班级的所有个体代表的特征子集放入分类器中进行训练,得到新班级中每个个体的fitness,与初始班级的对应个体fitness值进行比较. 若有所提升,则保留新个体;否则仍保留旧个体.

学生在上个阶段向最优个体(教师)移动后,整个班级水平提高到一定值,进入下一个进化阶段. 在学阶段,整个班级没有教师,所有个体都随机选择一位同伴互动,将自身水平提高至一定程度,并将更新后个体与原个体进行比较,择优保留. 因此,候选解决方案有一定概率朝着最优解决方案发展,整个过程可描述为

$$p^{t+1} = \begin{cases} p^t + r \times \left(p^t - q^t\right), f\left(p^t\right) > f\left(q^t\right) \\ p^t + r \times \left(q^t - p^t\right), f\left(q^t\right) > f\left(p^t\right) \end{cases} \tag{4}$$

观察这个模型可得,在教阶段,每个个体都向教师(最优个体)学习靠拢,搜索速度快,但种群多样性容易过早丢失,进而陷入局部最优. 在学阶段,通过个体在小范围内进行的相互学习,不会过早向全局最优方向聚集,因此该算法可保证种群多样性和全局搜索能力. 在进化过程中,生成的每个维度的数值与阈值δ进行比较,得到新的二进制向量. 另外,需要注意的是,整个进化过程中,变量的上界和下界一直在规范每个维度的数值,保证其在一定范围内.

3 局部搜索

3.1 局部搜索策略

局部搜索(Local Search,LS)是一种将搜索范围更改为接近最佳解决方案来扩大搜索范围,并期望找到更好解决方案的方法. 邻域即给定点附近其他点的集合,在特征选择问题中,邻域可被认为由给定操作对给

定问题域上的解进行扰动所得问题域上其他解的集合. 局部搜索操作从某个初始解开始,通过不同的邻域动作产生初始解的邻域解,根据某个指标或某种策略选择邻域解中优于初始解的个体,重复上述过程直到满足终止条件. 不同局部搜索算法区别在于邻域动作的定义以及选择邻域解的策略,这也是决定算法好坏的关键之处. 在本文中,采用的轻干扰方法包括三个邻域动作寻找新的邻域解,并采用轮盘赌的形式决定选择何种类型操作进行干扰. 这三个操作的实现细节可以描述如下:

(1)交换(Swap)操作

该操作在表示特征的二进制向量中,根据向量长度即特征个数,随机选取两个位置,将两个位置的元素进行交换得到一个新的特征向量,即原向量的邻域解.

(2)插入(Insert)操作

同交换操作一样,随机选择特征向量中两个位置,根据这两个位置下标值进行插入操作. 将下标较小位置上元素插入到下标较大位置,将较小位置后一位至较大位置之间的元素一并插入到空出来的位置区域.

(3)翻转(Reverse)操作

随机选择特征向量中两个位置,将这两个位置范围内的所有元素进行翻转,得到新的特征向量. 若随机选取的两个位置下标相差较小,则扰动范围较小;若两个位置下标相差较大,则扰动范围较大.

例如,假设对一个八维的二进制向量$\boldsymbol{v}$=(1 0 1 1 1 0 0 0),随机选取两个位置的下标分别为1和6,图3中(a)、(b)、(c)分别对应Swap,Insert,Reverse操作.

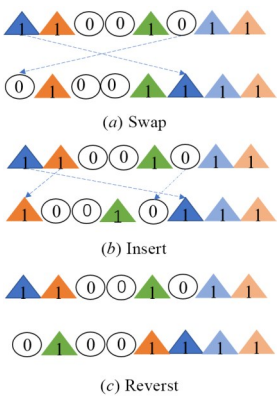

图3　三种扰动操作具体实现过程

在扰动过程中,为进一步增强搜索随机性,将交换扰动概率设置为0.3,插入扰动概率设置为0.2,翻转扰动概率设置为0.5,采用轮盘赌形式随机选择一种扰动. 第一种扰动操作是面向特征向量针对点对点的交换操作,扰动幅度较小,涉及范围也较小;第二种和第三种都是针对线对线的操作,扰动幅度较大,涉及范围也较广. 初始解在经过扰动得到邻域解后,若邻域解优于初始解,则邻域解代替初始解参加下一次迭代,否则初始解参与下一次迭代,直至满足结束条件为止,具体可见算法1. 然而,这些局部搜索操作也存在一些问题,例如:针对交换操作,若随机选取的两个位置上的元素值相等(均为0或均为1)则相当于原向量并没有发生变化. 针对插入操作,若随机选取的两个位置相距较近则扰动幅度较小. 这些局限都使得算法出现扰动失败的情况. 因此,本文将提出使用变邻域搜索机制对算法进行进一步优化.

算法1　局部搜索

输入: 向量$\boldsymbol{X}^{\text{old}}$, 迭代次数$N$.
输出: 向量$\boldsymbol{X}^{\text{new}}$.

```
将Swap, Insert, Reverse操作的概率分别设置为0.2, 0.3, 0.5
将n设置为输入向量的长度
在(1,n)随机选择两个位置i1,i2
FOR i=1:N
  在p(Swap), p(Insert), p(Reverse)中轮盘赌选择m
  IF m=0.2, THEN
    交换i1,i2位置上的元素,生成x'
  ELSE IF m=0.3
    将i2位置上的元素放在i1位置,i1与i2-1位置中的元素放入i1+1与i2位置之间,生成x'
  ELSE
    整体翻转i1与i2位置之间的所有元素,生成x'
  END IF
END FOR
IF fitness(x')>fitness(x), THEN x=x'
END IF
X^new=x
```

3.2　变邻域干扰策略

变邻域搜索(Variable Neighborhood Search,VNS),直接对整个可行解进行干扰,加强算法局部搜索能力. 该算法是一种改进型局部搜索算法,利用不同动作构成不同邻域交替搜索,在集中性和疏散性之间达到平衡. 相对于上一小节的局部搜索,VNS对整个特征向量进行扰动,范围更大,扰动力度更强. VNS基于以下三个规则:

(1)当前邻域局部最优解未必是其他邻域最优解;

(2)全局最优解是所有邻域局部最优解;

(3)局部最优解往往彼此相距很近.

由此可见,局部最优解往往会包含全局最优解的某些信息,则对局部最优解进行干扰,极有可能在其邻域中找到一个更优解. VNS一般包括两个主要部分:Shaking和VND(Variable Neighborhood Descent). 其中,Shaking同局部搜索的扰动操作类似,是一种扰动算子,

它负责扰动产生邻域解. 在VND中,若当前邻域找不到更优解时,则跳到下一个邻域搜索,如图蓝色实线;若在本邻域找到更优解则返回第一个邻域重新搜索,如图橙色虚线所示. 前者搜索全局最优解,后者验证找到的最优解是否为目前最优. 以初始向量$\boldsymbol{V}$=(1,1,0,0,1,0,0,0)为例,图4展示对其实施VNS策略的过程.

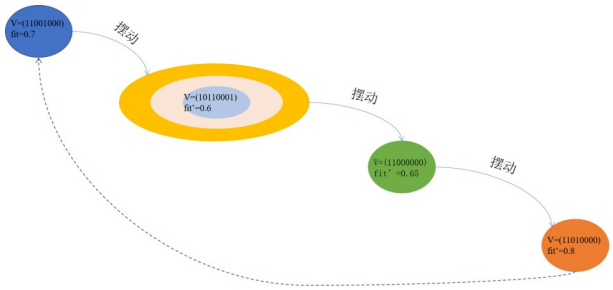

图4 VNS搜索最优解过程

整个VNS搜索过程伪代码可描述为算法2.

算法2	变邻域搜索
输入:	向量$\boldsymbol{X}^{\text{old}}$.
输出:	向量$\boldsymbol{X}^{\text{new}}$, 邻域数$k$.
1.	设置k=1
2.	IF fitness($\boldsymbol{X}^{\text{new}}$)>fitness($\boldsymbol{X}^{\text{old}}$),
3.	移动,且令$\boldsymbol{X}^{\text{old}}$=$\boldsymbol{X}^{\text{new}}$
4.	k=1
5.	ELSE
6.	k=k+1
7.	END IF
8.	$\boldsymbol{X}^{\text{new}}$=$\boldsymbol{X}^{\text{old}}$
9.	输出$\boldsymbol{X}^{\text{new}}$

4 BTLBOLS算法

从机器学习角度出发,由于癌症数据集基因数目庞大,在训练过程中总会出现过拟合及维度灾难,并对分类产生负面影响. 处理实际问题时,冗余特征会增加更多噪声对分类精度产生不利影响. 本研究引入一个新的Wrapper算法,通过TLBO和局部搜索的糅合,对它们进行互补. 在分类精度方面,需寻找更优质量特征子集的同时尽可能减少特征的数量. 此外,综合考虑特征数目以及分类精度两个冲突的指标,新的适应性函数也在此基础上进行改进. 所提出的方法有以下几个方面,如预处理、优化和分类. 作为数据处理的第一步,用Filter方法进行初步筛选,每一个基因都可以根据信息论进行估计并排序,然后创建一个维度较低的数据集,进而降低后续Wrapper方法的计算成本.

4.1 基因优化

Filter方法以有效降低数据维度的形式提供有意义的基因集合,减少基因表达向量. 然而该阶段对数据进行缩减所得结果仍存留噪声和冗余信息,无法从中获取相关的统计数据. 为降低入选基因的数量,获取信息量最大的基因与最高分类精度,本文提出新的混合Wrapper方法,称为(Binary Teaching and Learning-based Optimization-Local Search Algorithm,BTLBOLS),对KNN模型实现最大分类精度,估计最佳基因子集的预测适配度并降低过拟合风险. 在第一节框架基础上,我们将局部搜索策略与元启发式算法TLBO结合起来,从高维数据集中判别样本并估计分类精度.

为平衡算法开发与探索的能力,在算法初期,所有种群均被随机初始化,每个学习者均可被看作一个候选解,初始化后,计算每个个体的fitness函数值,并选取最大值对应的代理作为教学者,通过公式选取更新优化后最终符合要求的学习者作为局部搜索的初始解,搜索其邻域以期获取更优解. 在本方案中,每个学习者均以N维实数字符串形式存在,通过二进制编码方式转化为二进制向量进入分类器进行训练,整个过程重复十次,并在每个数据集中选择不同的进化基因,算法3给出所提出算法的伪代码.

4.2 BTLBOLS

最初,TLBOLS是为解决复杂非线性全局优化问题而发展起来的. 在最初的TLBOLS方法中,个体具有连续域的位置向量,在搜索空间中不断移动. 为要求搜索代理适应离散化的特征空间,我们提出了TLBOLS的二进制变体,称为BTLBOLS. 混合型BTLBOLS以LS作为局部优化器,开发一个高效基因选择算法,在探索和开发之间寻求更好的平衡. 此外,一个合适的评价函数提供一个概率调整各代理位置,以实现最终显著效益.

如图5所示,初始化阶段将连续的小数编码转化为非连续的二进制编码,生成一定数量的个体构成班级,将所有个体代表的特征子集放入分类器进行训练,得到每个个体的fitness. 第一个While循环表示当目前迭代次数小于最大迭代次数时,不断以下进化过程:教阶段提高班级整体平均水平,学阶段每个学生与其他学生交流提升自身水平;在结束教与学算法迭代后,生成一个最优个体,此时引入局部搜索和变邻域搜索机制进行干扰,局部搜索通过三种扰动操作对该个体进行小范围干扰,产生新的优秀个体;接着变邻域搜索通过改变最优个体的邻域,不断搜索更优解. 此外,为保证个体的质量,在整个进化过程和扰动过程中,每当产生新的个体,都要通过贪心选择与旧个体进行比较选择,保留最优解.

4.3 分类器

本文是基于K-Nearest Neighbor(KNN)分类器进行数据挖掘的. KNN分类器是一种基于K近邻算法的统计分类器. K近邻算法主要思想可概括为:近朱者赤,近墨者黑. KNN在解决分类问题时,针对一个数据集,输

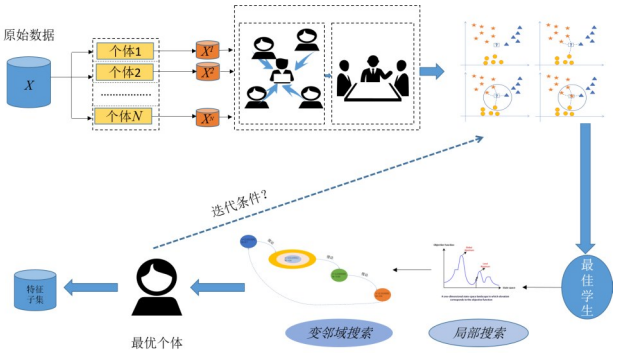

图5　BTLBO-LS流程图

入没有标签(标明数据类别的标记)即没有经过分类学习的新数据时,首先提取该新数据特征并与测试集数据(已分类)特征一一比较;然后从中选择K个与新数据最相近的数据标签,统计这K个最邻近数据中出现次数最多的分类,即少数服从多数,将其作为该新数据的类别. 当样本数足够多时,K越大,算法抗干扰能力越强,分类也越准确,可靠性也越高. 当样本数有限时,KNN算法的性能与K和度量样本距离的方式有着直接的关系. 本文采用欧氏距离度量样本点,假设有两个样本点$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$和$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)$,它们之间的欧氏距离为

$$\mathrm{d}(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}(x_i-y_i)^2 \tag{5}$$

4.4　评价指标

在特征选择过程中,为更全面对比不同算法性能,每个数据集均采用十折交叉验证. 将最终分类准确率Acc、精确率Precision、查全率Recall、F1分数F1-Score、AUC、目标函数值fitness、选取特征数Nfeat与初始维度D比值作为对比算法评价标准. 分类准确率Acc为

$$\mathrm{acc}=\frac{\mathrm{total}(\mathrm{predict\_label}=\mathrm{real\_label})}{\mathrm{total\_samples}} \tag{6}$$

$$\mathrm{Acc}=\frac{1}{N}\sum_{i=1}^{N}\mathrm{acc}_i \tag{7}$$

其中,acc为预测正确样本数占总体样本数比例;Acc为所有类别acc平均值. Nfeat即最优特征子集中特征的总数,定义如式(8),其中**fea_v**代表最优特征子集的二进制特征向量.

$$\mathrm{Nfeat}=\mathrm{total}(\mathbf{fea\_v}=1) \tag{8}$$

Precision,Recall,F1-Score,AUC这些性能指标根据式(9)~式(12)进行定义:

$$\mathrm{Precision}=\frac{\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i}}{N} \tag{9}$$

$$\mathrm{Recall}=\frac{\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i}}{N} \tag{10}$$

算法3　BTLBOLS

输入: Npop,　D, $t_{\max}$ L_k ,U_k, MaxIt, MaxSubIt.

输出: 向量$\boldsymbol{X}_{\mathrm{new}}$.

FOR i=1:Npop
　按照 $\mathrm{pop}_i^{\mathrm{old}}=L_k+\mathrm{rand}\times(U_k-L_k)$初始化个体并评价
END FOR
选出最优个体作为教师;
选出最差个体重新初始化
$\mathrm{pop}_{\mathrm{mean}}=\left(\sum_{i=1}^{n\mathrm{Pop}}\mathrm{pop}_i\right)\Big/\mathrm{Npop}$
WHILE $t\leqslant t_{\max}$
　t=t+1
　FOR i=1 to Npop
　　IF 该个体为教师, THEN TF=1
　　ELSE　$\mathrm{TF}=\dfrac{f(i)}{f(\mathrm{teacher})-f(i)}$
　　END IF
　END FOR
　FOR k=1 to D
　　$\mathrm{Diff}=\mathrm{pop}_{\mathrm{teacher}}-\mathrm{TF}\times\mathrm{pop}_{\mathrm{mean}}$
　　根据 $\mathrm{pop}_{i,k}^{new}=\mathrm{pop}_{i,k}^{\mathrm{old}}+\mathrm{rand}\times\mathrm{Diff}(k)$更新班级并评价
　END FOR
　IF fitness(新个体)>fitness(旧个体), THEN
　新个体代替旧个体
　ENF IF
　FOR i=1 to Npop
　随机选择同伴学习者 pop_p
　IF fitness(同伴)>fitness(该个体), THEN
　　按照 $\mathrm{pop}_i=\mathrm{pop}_i+\mathrm{rand}(\mathrm{pop}_i-\mathrm{pop}_p)$更新该个体
　ELSE
　　按照 $\mathrm{pop}_i=\mathrm{pop}_i+\mathrm{rand}(\mathrm{pop}_p-\mathrm{pop}_i)$更新该个体
　END IF
　END FOR
END WHILE
保存最优个体及其适应度值
WHILE j<MaxIt
　WHILE r<MaxSubIt
　　对上一步存储的最优个体局部搜索
　END WHILE
　保留最优个体
　WHILE k< $K_{\max}$
　　对上一步的最优个体进行变邻域搜索
　END WHILE
END WHILE
保留最优特征子集$\boldsymbol{X}_{\mathrm{new}}$及其适应度

$$\text{F1-Score}=\frac{\sum_{i=1}^{N}\frac{2*\text{TP}_i}{2*\text{TP}_i+\text{FP}_i+\text{FN}_i}}{N} \tag{11}$$

$$\text{AUC}=\frac{\sum \text{pred}_{\text{pos}}>\text{pred}_{\text{neg}}}{\text{posnumber}*\text{negnumber}} \tag{12}$$

TP,TN,FP,FN分别为True positive,True negative,False positive,False negative,即真正类、真负类、假正类、假负类,N为类别数. Precision着重评估所有预测为正例样本中实际正例样本数所占比例,Recall表示所有实际正例样本中预测为正例样本数所占比例,F1-Score将Precision和Recall这两个分值合并为一个分值,在合并过程中,召回率和精准率同等重要,这三个参数都是在样本类别不平衡的情况下Acc无法精准衡量算法优劣时采用的综合评价标准.

AUC全称为AreaUnderRoc,即ROC(Receiver Operating Characteristic)曲线下的面积,从统计学的角度来看,AUC为随机挑选一个正样本和负样本时,分类模型对正样本预测分数大于负样本预测分数概率,用于衡量整体样本间排序能力.

fitness为本文的目标函数,综合考虑Acc和Nfeat,其计算方式为

$$\text{fitness}(\boldsymbol{x})=\alpha\times\frac{\beta}{\theta}+(1-\alpha)\times\gamma \tag{13}$$

其中,fitness($\boldsymbol{x}$)中 $\boldsymbol{x}$表示特征向量,它可以描述特征子集分类能力,具体情况可见5.3.1节. 其中α为(0,0.1)之间的常量,本文取值0.03(具体讨论见5.3.1节),β为特征向量长度,θ为初始特征向量长度400(具体讨论见5.3.1节),γ为分类器准确率.

与使用准确率和特征数量这两个目标相比,使用以上七个目标可以更充分地解释癌症基因表达数据.

5 实验结果分析

在实验中,算法使用MATLAB编程语言编写,所有实验均内存为16GB、操作系统为Windows10的PC机上执行.

5.1 数据集

在本节中,我们在10个基因表达数据集上评估了所提出的方法和其他Wrapper方法. 关于这些数据集的更详细的信息可在www.gems-system.org获得. 表1总结了10个癌症基因表达数据集的主要信息,包括特征、类别和样本的数量. 样本数量从60到203不等,类别数从2到11不等,它们都包含成千上万个特征.

其中ColonTumor,DLBCL和Prostate_cancer有两个类别,其余数据集含有多个类别. 首先,通过互信息方法将数据集降维至400,然后采用十折交叉验证法将数据分为训练集和测试集. 训练集用于构建模型,并在进化过程中对个体进行评估;测试集对训练集训练的模型进行测试.

表1 10个癌症数据集的主要信息

序号	数据集	基因	样本	类别
1	9_Tumors	5 726	60	9
2	11_Tumors	12 533	174	4
3	Leukemia1	5 327	72	3
4	Leukemia2	11 225	72	3
5	Brain_Tumor1	5 920	90	5
6	ColonTumor	2 000	62	2
7	Prostate_cancer	10 509	102	2
8	DLBCL	5 469	77	2
9	SBRCT	2 308	83	4
10	Lung_cancer	12 600	174	11

5.2 实验参数设置

为保证实验公平性,加入扰动后算法与原始算法均采用相同参数;为保证实验结果稳定性,所有算法均运行10次. 各个参数具体设置如表2所示:其中,在特征选择过程中,T为TLBO算法迭代次数,Npop为班级中个体数,MaxIt是最大迭代次数,每次包含5次MaxSubIt迭代. K_1为VNS中寻求邻域个数最大值,K_2为KNN分类器中选取的最近邻个数.

表2 BTLBOLS算法参数信息

序号	参数	值
1	Npop(初始个体数)	100
2	R(算法运行次数)	10
3	T(迭代次数)	50
4	D(特征向量长度)	400
5	MaxIt(一级迭代次数)	20
6	MaxSubIt(二级迭代次数)	5
7	K_1(VNS最大邻域数)	10
8	K_2(KNN参数)	8
9	K-fold(交叉验证参数)	10

实验中评价指标均采用Acc,实验结果如图6所示. 其中9_Tumors(D) - SRBCT(D)是针对特征向量长度的实验,由图可得,当D=400时,算法表现突出,且此时10个数据集Acc均值达到最高;9_Tumors(T) - SRBCT(T)针对算法迭代次数进行对比,由图可知,当T=50时,算法可在特征选择问题上发挥最大优势;9_Tumors(Npop) - SRBCT(Npop)是针对进化种群中的初始个体进行实验,由图可得,当Npop=100时,算法在10个数据集上表现突出并取得最大Acc.

由上可得算法迭代次数T设置为50,班级中的个体数Npop设置为100,特征向量长度D为400时,算法效果最佳. 此外,我们对评价函数中的∂进行取值实验,从0.01开始,以0.02为步长,设计了4组实验,具体实验结果可如表3所示.

图6　不同参数的讨论分析结果

由表3可以看出,当∂取值为0.01时,在Brain_Tumor1,DLBCL,Leukemia2,Lung_Cancer这四个数据集上表现最优;取值为0.03时,在9_Tumors,ColonTumor,Leukemia1,Prostate_Tumor这四个数据集上表现优异.此外,当∂取0.05时,算法在11_Tumors表现优异,无论取何值,算法在SRBCT数据集上的表现都比较稳定且突出.当∂取值为0.03时,算法在六个数据集上所得Precision优于0.01时的情况,因此本文选用0.03作为最终参数.同时,将∂取值为0.03时fitness的变化趋势展示如下,可见采用该参数设置的算法性能稳定,适应度最终都趋于收敛.在此基础上,采用相同参数运行BTLBOLS算法,结果如表4所示,其中最优评价指标及对应的算法已加粗标出,在10个数据集上,BTLBO算法随着迭代次数增加的适应度函数的变化结果显示在图7.

由表4可直观地看出:对于11_Tumors,ColonTumor,DLBCL,Leukemia1,Lung_Cancer,BTLBO算法加入扰动后的结果相较于BTLBO有所提高,其原因可能是在数据集特征较多、类别数也较多的情况下BTLBO算法以较大概率去除了一些高效特征进行分类学习,在加入扰动后,增强了算法的局部搜索能力,使得算法不再受囿于局部最优解;对于9_Tumors,Brain_Tumor1,Leukemia2,Prostate_Tumor和SRBCT,BTLBO算法和BTLBOLS算法的表现基本持平,其原因可能是BTLBO

表3 ∂实验

数据集	∂	fitness	Acc	Nfeat/D	Precision	F1-score	AUC	Recall
9_Tumors	0.01	0.678 8	0.680 0	0.564 7	0.590 1	0.511 1	0.840 7	0.624 1
	0.03	**0.677 6**	**0.683 3**	**0.492 5**	**0.593 5**	**0.514 4**	**0.828 5**	**0.629 0**
	0.05	0.679 3	0.686 7	0.539 7	0.639 4	0.544 6	0.832 1	0.632 4
	0.07	0.666 8	0.676 7	0.535 5	0.613 3	0.524 0	0.824 8	0.623 3
11_Tumors	0.01	0.846 9	0.852 1	0.330 2	0.848 9	0.766 3	0.896 5	0.761 0
	0.03	0.840 4	0.851 3	0.488 0	0.897 1	0.794 0	0.893 2	0.765 7
	0.05	**0.834 7**	**0.852 8**	**0.490 3**	**0.890 4**	**0.793 2**	**0.893 1**	**0.772 3**
	0.07	0.828 6	0.847 2	0.581 5	0.874 0	0.772 7	0.882 6	0.758 1
Brain_Tumor	**0.01**	**0.863 2**	**0.868 9**	**0.299 3**	**0.631 5**	**0.476 1**	**0.754 9**	**0.572 0**
	0.03	0.852 1	0.866 7	0.379 5	0.594 5	0.447 3	0.728 2	0.565 7
	0.05	0.845 2	0.866 7	0.437 0	0.629 0	0.476 4	0.752 3	0.571 3
	0.07	0.836 6	0.852 2	0.629 8	0.557 9	0.425 8	0.739 2	0.544 7
ColonTumor	0.01	0.907 4	0.912 9	0.363 0	0.918 7	0.899 0	0.849 1	0.885 0
	0.03	**0.899 4**	**0.913 6**	**0.439 7**	**0.919 1**	**0.898 8**	**0.868 4**	**0.884 0**
	0.05	0.883 1	0.906 0	0.449 5	0.914 0	0.891 2	0.835 7	0.874 9
	0.07	0.879 4	0.902 1	0.577 0	0.909 0	0.887 6	0.844 6	0.872 4
DLBCL	**0.01**	**0.978 2**	**0. 984 8**	**0.327 5**	**0.975 7**	**0.979 8**	**0.893 6**	**0.984 3**
	0.03	0.965 0	0.983 4	0.369 5	0.969 9	0.978 3	0.895 6	0.987 0
	0.05	0.954 2	0.970 7	0.641 2	0.949 0	0.962 4	0.897 5	0.976 6
	0.07	0.948 2	0.962 9	0.754 0	0.934 1	0.916 3	0.894 9	0.840 7
Leukemia1	0.01	0.991 0	0.997 5	0.344 5	0.997 9	0.996 3	0.998 1	0.995 0
	0.03	**0.980 2**	**0.998 6**	**0.386 0**	**0.999 1**	**0.998 9**	**0.979 2**	**0.998 6**
	0.05	0.967 0	0.993 4	0.466 2	0.992 8	0.992 0	0.992 4	0.991 4
	0.07	0.959 4	0.992 1	0.524 8	0.994 4	0.990 6	0.983 8	0.987 3
Leukemia2	**0.01**	**0.983 3**	**0.989 8**	**0.334 2**	**0.988 9**	**0.988 4**	**0.993 0**	**0.988 0**
	0.03	0.970 3	0.988 2	0.390 0	0.987 5	0.986 6	0.992 1	0.986 0
	0.05	0.961 7	0.983 0	0.556 5	0.983 7	0.981 0	0.985 6	0.978 9
	0.07	0.953 2	0.983 0	0.556 7	0.985 0	0.982 0	0.983 6	0.979 4
Lung_Cancer	**0.01**	**0.955 2**	**0.961 0**	**0.377 0**	**0.953 6**	**0.953 3**	**0.954 1**	**0.953 3**
	0.03	0.942 9	0.955 8	0.526 2	0.960 7	0.922 8	0.944 1	0.893 4
	0.05	0.936 5	0.956 3	0.560 0	0.970 9	0.927 8	0.925 0	0.894 6
	0.07	0.928 0	0.950 5	0.629 0	0.951 6	0.898 0	0.922 2	0.864 7
Prostate_Tumor	0.01	0.945 3	0.953 6	0.117 0	0.953 6	0.953 3	0.954 1	0.953 2
	0.03	**0.931 9**	**0.955 1**	**0.183 2**	**0.955 9**	**0.955 4**	**0.958 0**	**0.955 3**
	0.05	0.916 7	0.943 3	0.411 5	0.943 6	0.943 4	0.930 7	0.943 5
	0.07	0.904 0	0.938 7	0.442 2	0.938 9	0.938 6	0.942 4	0.938 6
SRBCT	**0.01**	**0.997 0**	**1.000 0**	**0.703 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	0.03	**0.991 0**	**1.000 0**	**0.701 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	0.05	**0.984 1**	**1.000 0**	**0.681 5**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	0.07	**0.979 7**	**1.000 0**	**0.710 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**

算法在这四个数据集上已表现出优异的准确率,加入扰动策略很可能在最优解附近扰动,所得结果与原始算法大致相同.

5.2.1 编码实验

最初的教与学优化(TLBO)算法是基于求解连续空间复杂函数优化问题被提出的,本文使用编码转换方法,将其设计成能求解特征选择问题的二进制TLBO算法(BTLBO). 在这一节中,对两种编码方式分别进行探讨.

小数编码:用一个二元组($\boldsymbol{X}$, $\boldsymbol{Y}$)表示每个学生个体或临时个体. 实向量 $\boldsymbol{X}=[x_1, x_2, \cdots, x_m]$, $x_i\in[0, 1]$, $i=$

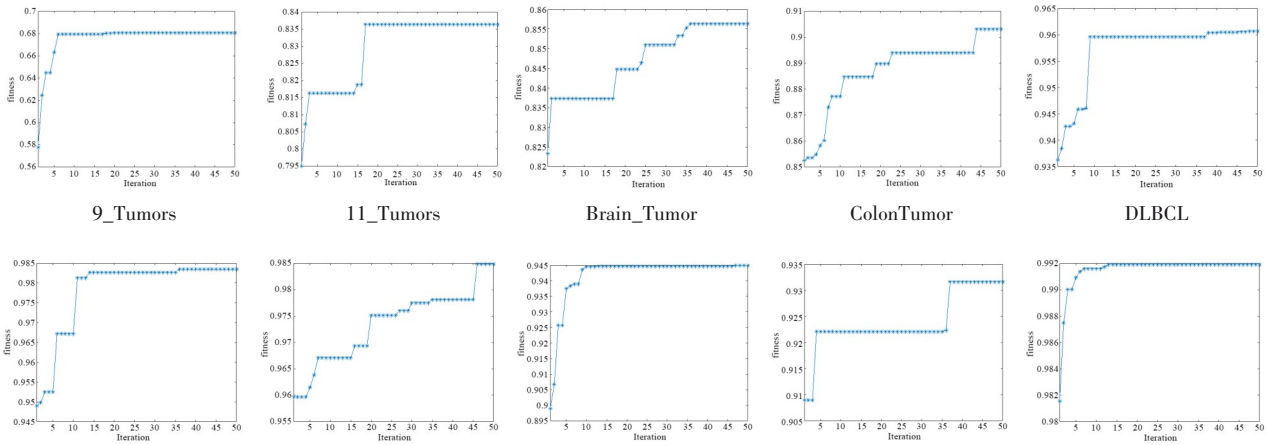

图7　在10个数据集上BTLBO算法随着迭代次数增加的适应度函数的变化结果

表4　BTLBO和BTLBOLS算法对比

数据集	算法	Nfeat/D	fitness	Acc	Precision	F1-score	AUC	Recall
9_Tumors	BTLBO	0.492 5	0.677 6	0.683 3	0.593 5	0.514 4	0.828 5	0.629 0
	BTLBOLS	0.493 5	0.677 6	0.683 3	0.609 7	0.520 2	0.822 9	0.628 0
11_Tumors	BTLBO	0.488 0	0.840 4	0.851 3	0.897 1	0.794 0	0.893 2	0.765 7
	BTLBOLS	**0.437 0**	**0.848 7**	**0.861 5**	**0.875 8**	**0.791 4**	**0.902 8**	**0.776 6**
Brain_Tumor1	**BTLBO**	**0.379 5**	**0.852 1**	**0.866 7**	**0.594 5**	**0.447 3**	**0.728 2**	**0.565 7**
	BTLBOLS	0.531 3	0.849 1	0.858 9	0.572 7	0.428 6	0.730 6	0.548 7
ColonTumor	BTLBO	0.439 7	0.899 4	0.913 6	0.919 1	0.898 8	0.868 4	0.884 0
	BTLBOLS	**0.319 0**	**0.901 5**	**0.919 5**	**0.929 2**	**0.908 0**	**0.855 4**	**0.892 3**
DLBCL	BTLBO	0.369 5	0.965 0	0.983 4	0.969 9	0.978 3	0.895 6	0.987 0
	BTLBOLS	**0.311 8**	**0.967 2**	**0.987 5**	**0.980 0**	**0.982 9**	**0.908 0**	**0.986 0**
Leukemia1	BTLBO	0.386 0	0.980 2	0.998 6	0.999 1	0.998 9	0.979 2	0.998 7
	BTLBOLS	**0.432 5**	**0.971 6**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
Leukemia2	**BTLBO**	**0.390 0**	**0.970 3**	**0.988 2**	**0.987 5**	**0.986 6**	**0.992 1**	**0.986 0**
	BTLBOLS	0.400 0	0.956 4	0.985 7	0.986 7	0.984 9	0.991 7	0.983 3
Lung_Cancer	BTLBO	0.526 2	0.942 9	0.955 8	0.960 7	0.922 8	0.944 1	0.893 4
	BTLBOLS	**0.552 5**	**0.945 3**	**0.966 0**	**0.979 9**	**0.944 7**	**0.910 5**	**0.915 4**
Prostate_Tumor	**BTLBO**	**0.183 2**	**0.931 9**	**0.955 1**	**0.955 9**	**0.955 4**	**0.958 0**	**0.955 4**
	BTLBOLS	0.185 0	0.912 6	0.950 9	0.952 1	0.951 6	0.952 9	0.951 5
SRBCT	**BTLBO**	**0.701 0**	**0.991 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	BTLBOLS	**0.750 0**	**0.987 5**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**

1, 2, …, m, 由随机初始化生成,同时有一个与之对应的二进制向量 $\boldsymbol{Y}=[y_1, y_2, \cdots, y_m]$, $y_i\in\{0, 1\}$, i=1, 2, …, m, 二进制向量利用从连续空间[0,1]到离散空间{0,1}的映射得到,映射关系如式(14)所示. 算法中个体的进化针对个体的实向量 $\boldsymbol{X}$ 进行,二进制向量 $\boldsymbol{Y}$ 按式随之产生,个体的适应度由二进制向量 $\boldsymbol{Y}$ 计算得到.

$$y_i=\begin{cases}1, & x_i>0.5\\ 0, & x_i\leqslant 0.5\end{cases} \tag{14}$$

二进制编码:用一个二元组($\boldsymbol{Y}$, $\boldsymbol{X}$)表示每个学生个体或临时个体. 二进制向量 $\boldsymbol{Y}=[y_1, y_2, \cdots, y_m]$, $y_i\in\{0, 1\}$, i=1, 2, …, m, 同时有一个与之对应的实向量 $\boldsymbol{X}=[x_1, x_2, \cdots, x_m]$, $x_i\in[0, 1]$, i=1, 2, …, m, 二进制向量由随机初始化的二进制编码得到,实向量利用从离散空间{0,1}到连续空间[0,1]到的映射得到,映射关系如式(15)所示. 算法中个体的进化针对个体的实向量 $\boldsymbol{X}$ 进行,在计算个体适应度时,进化过程中的实向量由式(14)映射到二进制向量 $\boldsymbol{Y}$. 两种编码方式的比较结果如表5所示,最优编码方式已用粗体标出,可以看出采

表5 不同编码方式对比结果

数据集	编码	Nfeat/D	fitness	Acc	Precision	F1-score	AUC	Recall
9_Tumors	Binary	0.492 5	0.629 1	0.633 3	0.594 3	0.507 6	0.844 7	0.591 7
	Continuous	**0.492 5**	**0.677 6**	**0.683 3**	**0.593 5**	**0.514 4**	**0.828 5**	**0.629 0**
11_Tumors	Binary	0.507 5	0.801 7	0.810 8	0.870 8	0.738 9	0.881 1	0.698 7
	Continuous	**0.488 0**	**0.840 4**	**0.851 3**	**0.897 1**	**0.794 0**	**0.893 2**	**0.765 7**
Brain_Tumor1	**Binary**	**0.175 0**	**0.856 7**	**0.877 8**	**0.551 4**	**0.421 1**	**0.743 4**	**0.580 0**
	Continuous	0.379 5	0.852 1	0.866 7	0.594 5	0.447 3	0.728 2	0.565 7
ColonTumor	Binary	0.157 5	0.882 3	0.904 8	0.915 4	0.891 1	0.779 2	0.874 0
	Continuous	**0.439 7**	**0.899 4**	**0.913 6**	**0.919 1**	**0.898 8**	**0.868 4**	**0.884 0**
DLBCL	Binary	0.160 0	0.948 8	0.973 2	0.965 1	0.965 0	0.850 0	0.965 1
	Continuous	**0.369 5**	**0.965 0**	**0.983 4**	**0.969 9**	**0.978 3**	**0.895 6**	**0.987 0**
Leukemia1	Binary	0.180 0	0.963 3	0.987 5	0.991 5	0.989 0	0.987 5	0.986 7
	Continuous	**0.386 0**	**0.980 2**	**0.998 6**	**0.999 1**	**0.998 9**	**0.979 2**	**0.998 7**
Leukemia2	Binary	0.182 5	0.949 5	0.973 2	0.971 8	0.970 5	0.985 1	0.969 4
	Continuous	**0.390 0**	**0.970 3**	**0.988 2**	**0.987 5**	**0.986 6**	**0.992 1**	**0.986 0**
Lung_Cancer	Binary	0.547 5	0.934 5	0.946 4	0.974 0	0.917 2	0.914 4	0.875 0
	Continuous	**0.526 2**	**0.942 9**	**0.955 8**	**0.960 7**	**0.922 8**	**0.944 1**	**0.893 4**
Prostate_Tumor	Binary	0.140 0	0.916 0	0.940 0	0.941 5	0.941 4	0.944 6	0.941 5
	Continuous	**0.183 2**	**0.931 9**	**0.955 1**	**0.955 9**	**0.955 4**	**0.958 0**	**0.955 4**
SRBCT	Binary	0.545 0	0.986 3	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0
	Continuous	**0.701 0**	**0.991 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**

用连续的实数编码(Continuous)的算法表现明显优于采用二进制编码(Binary)的算法,故本文采用实数编码映射到离散编码的编码机制.

$$x_i=\begin{cases}\text{rand}(0,0.5), & y_i=0\\ \text{rand}(0.5,1), & y_i=1\end{cases} \tag{15}$$

5.2.2 对比实验

我们将本文提出的算法分别与其他经典特征选择算法、进化特征选择算法进行比较,详细信息总结如表6所示. 前五个为特征选择算法,其中CMIM(Conditional Mutual Information)、JMI(Joint Mutual Information)、MRMR(Minimal Redundancy Maximum Relevance)、DISR(Double Input Symmetrical Relevance)这四种特征选择方法基于条件互信息进行特征选择,Relief-F则基于相似性选择特征;后五个为新型进化特征选择算法,其中经典特征选择算法的详细信息可在https://jundongl.github.io/scikit-feature/algorithms.html获得,进化特征选择算法的代码可在https://github.com/JingweiToo/Wrapper-Feature-Selection-Toolbox获得.

表7和表8分别总结了特征选择算法、进化特征选择算法与本文算法对比实验结果,并用粗体标记最优算法. 从中得出,本文提出的算法在fitness,Acc,Precision,F1-score,AUC,Recall这六个指标上的表现都优于其他特征选择算法,在9_Tumors数据集上,其他特征选择算法可以得到很高的Precision却难以取得较好的Acc,其原因可能是该数据集样本极度不平衡,而本文所提算法在9_Tumors数据集上表现突出,结果稳定.

表6 对比算法

特征选择算法名称	进化特征选择算法名称
CMIM/2006[35]	MPA(Marine Predators Algorithm)/2020[40]
JMI/1999[36]	GNDO(Generalized Normal Distribution Optimization)/2020[41]
mRMR/2005[37]	SMA(Slime Mould Algorithm)/2020[42,43]
DISR/2008[38]	EO(Equilibrium Optimizer)/2020[44]
Relief-F/2003[39]	MRFO(Manta Ray Foraging Optimization)/2020[45]

5.3 补充实验

除基本实验外,我们在另外35个基因表达数据集上进行实验,同样获得较好的结果,进一步证实了本文算法的性能,为方便比较不同算法的分类效率,此处将实验结果总结在图8中. 从中可得出:BTLBOLS在Alizadeh-2000-v2,Armstrong-2002-v1,Liang-2005,Nutt-2003-v2,Nutt-2003-v3这五个数据集上Acc与BTLBO的Acc持平且处于最高水平,在另外30个数据集上BTLBOLS的Acc均高于BTLBO;在所有7个算法里,BTLBOLS在29个数据集上Acc达到最高水平,在Bhattacharjee-2001,Chen-2002和Ramaswamy-2001数据集上虽然没有达到最高水平,但也处于次高水平. 总体而言,本文提出的BTLBO与BTLBOLS效果稳定,与其他

表7　经典特征选择算法对比实验

数据集	算法	fitness	Acc	Precision	F1-score	AUC	Recall
9_Tumors	**BTLBO**	**0.677 6**	**0.683 3**	**0.593 5**	**0.514 4**	**0.828 5**	**0.629 0**
	BTLBOLS	**0.677 6**	**0.683 3**	**0.609 7**	**0.520 2**	**0.822 9**	**0.628 0**
	CMIM	0.228 4	0.233 3	0.901 7	0.144 4	0.564 4	0.200 8
	JMI	0.325 4	0.333 3	0.915 0	0.159 8	0.604 0	0.289 2
	mRMR	0.244 6	0.250 0	0.904 2	0.127 6	0.573 9	0.218 0
	DISR	0.228 4	0.233 3	0.902 4	0.131 3	0.572 0	0.194 4
	Relief-F	0.390 1	0.400 0	0.923 5	0.227 1	0.694 9	0.355 2
11_Tumors	BTLBO	0.840 4	0.851 3	0.897 1	0.794 0	0.893 2	0.765 7
	BTLBOLS	**0.848 7**	**0.861 5**	**0.875 8**	**0.791 4**	**0.902 8**	**0.776 6**
	CMIM	0.642 6	0.661 4	0.965 2	0.572 6	0.752 7	0.556 1
	JMI	0.685 3	0.705 6	0.969 5	0.573 4	0.819 1	0.597 8
	mRMR	0.362 3	0.372 5	0.934 6	0.289 2	0.641 1	0.337 4
	DISR	0.665 4	0.685 0	0.967 0	0.576 1	0.786 2	0.584 3
	Relief-F	0.758 3	0.780 7	0.977 1	0.752 3	0.866 1	0.702 3
Brain_Tumor1	**BTLBO**	**0.852 1**	**0.866 7**	**0.594 5**	**0.447 3**	**0.728 2**	**0.565 7**
	BTLBOLS	0.849 1	0.858 9	0.572 7	0.428 6	0.730 6	0.548 7
	CMIM	0.724 1	0.744 4	0.846 7	0.310 6	0.601 7	0.340 0
	JMI	0.648 7	0.666 7	0.804 2	0.047 5	0.451 3	0.200 0
	mRMR	0.648 7	0.666 7	0.800 0	0.047 1	0.500 0	0.200 0
	DISR	0.659 5	0.677 8	0.810 8	0.104 6	0.528 3	0.220 0
	Relief-F	0.756 5	0.777 8	0.875 0	0.340 1	0.607 0	0.416 7
ColonTumor	BTLBO	0.899 4	0.913 6	0.919 1	0.898 8	0.868 4	0.884 0
	BTLBOLS	**0.901 5**	**0.919 5**	**0.929 2**	**0.908 0**	**0.855 4**	**0.892 3**
	CMIM	0.705 8	0.721 4	0.634 1	0.641 5	0.593 3	0.634 1
	JMI	0.745 0	0.761 9	0.679 5	0.697 4	0.597 5	0.679 5
	mRMR	0.708 1	0.723 8	0.634 1	0.641 5	0.637 5	0.634 1
	DISR	0.722 0	0.738 1	0.656 8	0.670 4	0.587 5	0.656 8
	Relief-F	0.735 8	0.752 4	0.679 5	0.697 4	0.587 0	0.679 5
DLBCL	BTLBO	0.965 0	0.983 4	0.969 9	0.978 3	0.895 6	0.987 0
	BTLBOLS	**0.967 2**	**0.987 5**	**0.980 0**	**0.982 9**	**0.908 0**	**0.986 0**
	CMIM	0.821 5	0.844 6	0.755 0	0.767 2	0.711 1	0.755 0
	JMI	0.870 0	0.894 6	0.824 9	0.847 6	0.783 3	0.824 9
	mRMR	0.721 0	0.741 1	0.491 4	0.271 4	0.491 7	0.491 4
	DISR	0.745 3	0.766 1	0.544 0	0.506 9	0.533 3	0.544 0
	Relief-F	0.922 0	0.948 2	0.947 8	0.933 5	0.950 0	0.947 8
Leukemia1	BTLBO	0.980 2	0.998 6	0.999 1	0.998 9	0.979 2	0.998 7
	BTLBOLS	**0.971 6**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	CMIM	0.700 3	0.719 6	0.817 5	0.588 8	0.721 4	0.552 8
	JMI	0.715 9	0.735 7	0.821 9	0.719 8	0.753 3	0.661 0
	mRMR	0.516 7	0.530 4	0.696 5	0.271 3	0.567 8	0.374 4
	DISR	0.693 4	0.712 5	0.799 5	0.658 6	0.740 0	0.586 9
	Relief-F	0.785 2	0.807 1	0.862 7	0.767 3	0.808 6	0.694 8

表7 （续）

数据集	算法	fitness	Acc	Precision	F1-score	AUC	Recall
Leukemia2	**BTLBO**	**0.970 3**	**0.988 2**	**0.987 5**	**0.986 6**	**0.992 1**	**0.986 0**
	BTLBOLS	0.956 4	0.985 7	0.986 7	0.984 9	0.991 7	0.983 3
	CMIM	0.713 0	0.733 9	0.868 7	0.709 1	0.791 5	0.716 3
	JMI	0.764 9	0.787 5	0.895 0	0.766 2	0.837 5	0.769 0
	mRMR	0.553 6	0.569 6	0.780 0	0.522 7	0.654 7	0.540 9
	DISR	0.773 6	0.796 4	0.897 0	0.771 4	0.833 6	0.777 4
	Relief-F	0.836 0	0.860 7	0.929 7	0.857 5	0.901 1	0.855 2
Lung_Cancer	BTLBO	0.942 9	0.955 8	0.960 7	0.922 8	0.944 1	0.893 4
	BTLBOLS	**0.945 3**	**0.966 0**	**0.979 9**	**0.944 7**	**0.910 5**	**0.915 4**
	CMIM	0.728 5	0.750 0	0.857 0	0.506 3	0.617 6	0.440 1
	JMI	0.807 7	0.831 7	0.904 0	0.527 7	0.732 1	0.548 6
	mRMR	0.780 2	0.803 3	0.879 1	0.382 8	0.684 0	0.476 5
	DISR	0.827 3	0.851 9	0.916 5	0.564 4	0.716 5	0.575 3
	Relief-F	0.870 0	0.896 0	0.938 5	0.625 0	0.796 4	0.666 1
Prostate_Tumor	**BTLBO**	**0.931 9**	**0.955 1**	**0.955 9**	**0.955 4**	**0.958 0**	**0.955 4**
	BTLBOLS	0.912 6	0.950 9	0.952 1	0.951 6	0.952 9	0.951 5
	CMIM	0.734 8	0.756 4	0.756 2	0.756 5	0.719 8	0.756 2
	JMI	0.754 2	0.776 4	0.773 5	0.774 4	0.789 3	0.773 5
	mRMR	0.752 5	0.774 5	0.776 9	0.779 9	0.779 5	0.776 9
	DISR	0.781 6	0.804 5	0.803 1	0.803 9	0.812 7	0.803 1
	Relief-F	0.778 0	0.800 9	0.803 8	0.803 8	0.793 7	0.803 8
SRBCT	**BTLBO**	**0.991 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	BTLBOLS	**0.991 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	CMIM	0.802 8	0.822 2	0.937 5	0.832 9	0.895 3	0.828 4
	JMI	0.730 0	0.747 2	0.916 9	0.778 3	0.848 6	0.744 6
	mRMR	0.872 8	0.894 4	0.960 7	0.898 3	0.932 6	0.884 5
	DISR	0.810 8	0.830 6	0.941 3	0.840 7	0.885 8	0.827 0
	Relief-F	0.866 1	0.887 5	0.963 8	0.896 5	0.928 0	0.885 9

表8 进化特征选择算法对比实验

数据集	算法	fitness	Acc	Precision	F1-score	AUC	Recall
9_Tumors	**BTLBO**	**0.677 6**	**0.683 3**	**0.593 5**	**0.514 4**	**0.828 5**	**0.629 0**
	BTLBOLS	**0.677 6**	**0.683 3**	**0.609 7**	**0.520 2**	**0.822 9**	**0.628 0**
	MPA	0.307 6	0.316 7	0.912 4	0.167 8	0.629 7	0.285 9
	GNDO	0.418 4	0.416 7	0.925 8	0.239 6	0.657 0	0.363 8
	SMA	0.177 9	0.183 3	0.897 1	0.087 0	0.525 4	0.158 3
	EO	0.309 7	0.316 7	0.912 5	0.257 4	0.616 0	0.279 5
	MRFO	0.328 8	0.333 3	0.914 9	0.202 0	0.647 8	0.292 5
11_Tumors	BTLBO	0.840 4	0.851 3	0.897 1	0.794 0	0.893 2	0.765 7
	BTLBOLS	**0.848 7**	**0.861 5**	**0.875 8**	**0.791 4**	**0.902 8**	**0.776 6**
	MPA	0.703 3	0.723 5	0.971 3	0.627 5	0.810 6	0.640 8
	GNDO	0.739 1	0.747 1	0.973 7	0.658 7	0.869 0	0.645 8
	SMA	0.750 9	0.769 9	0.976 1	0.683 7	0.857 8	0.709 7
	EO	0.738 8	0.759 2	0.974 8	0.689 8	0.829 4	0.658 9
	MRFO	0.763 3	0.781 4	0.977 0	0.759 2	0.828 8	0.706 8

表 8 （续）

数据集	算法	fitness	Acc	Precision	F1-score	AUC	Recall
Brain_Tumor1	**BTLBO**	**0.852 1**	**0.866 7**	**0.594 5**	**0.447 3**	**0.728 2**	**0.565 7**
	BTLBOLS	0.849 1	0.858 9	0.572 7	0.428 6	0.730 6	0.548 7
	MPA	0.711 5	0.733 3	0.852 5	0.285 8	0.533 6	0.353 3
	GNDO	0.768 9	0.777 8	0.875 0	0.332 8	0.632 5	0.400 0
	SMA	0.743 7	0.766 7	0.885 0	0.293 9	0.666 8	0.413 3
	EO	0.798 6	0.822 2	0.922 5	0.495 7	0.698 9	0.526 7
	MRFO	0.689 8	0.711 1	0.851 7	0.171 1	0.540 7	0.313 3
ColonTumor	BTLBO	0.899 4	0.913 6	0.919 1	0.898 8	0.868 4	0.884 0
	BTLBOLS	**0.901 5**	**0.919 5**	**0.929 2**	**0.908 0**	**0.855 4**	**0.892 3**
	MPA	0.679 0	0.700 0	0.650 0	0.646 9	0.671 7	0.650 0
	GNDO	0.780 8	0.790 5	0.725 0	0.746 7	0.715 0	0.725 0
	SMA	0.706 7	0.728 6	0.685 2	0.685 0	0.652 5	0.685 2
	EO	0.764 5	0.788 1	0.735 2	0.750 4	0.719 2	0.735 2
	MRFO	0.795 9	0.819 0	0.780 7	0.794 3	0.725 0	0.780 7
DLBCL	BTLBO	0.965 0	0.983 4	0.969 9	0.978 3	0.895 6	0.987 0
	BTLBOLS	**0.967 2**	**0.987 5**	**0.980 0**	**0.982 9**	**0.908 0**	**0.986 0**
	MPA	0.883 4	0.910 7	0.886 6	0.879 5	0.910 7	0.886 6
	GNDO	0.909 3	0.923 2	0.877 5	0.890 2	0.907 4	0.877 5
	SMA	0.808 9	0.833 9	0.746 4	0.752 2	0.617 9	0.746 4
	EO	0.859 4	0.885 7	0.816 2	0.830 7	0.740 0	0.816 2
	MRFO	0.831 6	0.857 1	0.763 6	0.783 4	0.687 5	0.763 6
Leukemia1	BTLBO	0.980 2	0.998 6	0.999 1	0.998 9	0.979 2	0.998 7
	BTLBOLS	**0.971 6**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	MPA	0.744 8	0.767 9	0.880 3	0.441 3	0.786 2	0.587 4
	GNDO	0.717 7	0.725 0	0.806 6	0.597 5	0.737 5	0.548 3
	SMA	0.838 4	0.864 3	0.907 4	0.864 0	0.892 8	0.828 4
	EO	0.879 1	0.905 4	0.935 9	0.912 3	0.925 8	0.911 2
	MRFO	0.887 1	0.914 3	0.954 7	0.923 4	0.946 7	0.914 5
Leukemia2	**BTLBO**	**0.970 3**	**0.988 2**	**0.987 5**	**0.986 6**	**0.992 1**	**0.986 0**
	BTLBOLS	0.956 4	0.985 7	0.986 7	0.984 9	0.991 7	0.983 3
	MPA	0.847 0	0.873 2	0.940 0	0.881 7	0.917 5	0.880 1
	GNDO	0.877 2	0.889 3	0.944 8	0.878 1	0.909 2	0.878 9
	SMA	0.805 5	0.830 4	0.918 0	0.819 6	0.875 3	0.821 4
	EO	0.929 5	0.929 5	0.979 1	0.958 4	0.969 4	0.959 5
	MRFO	0.783 0	0.807 1	0.901 0	0.799 2	0.857 2	0.794 8
Lung_Cancer	BTLBO	0.942 9	0.955 8	0.960 7	0.922 8	0.944 1	0.893 4
	BTLBOLS	**0.945 3**	**0.966 0**	**0.979 9**	**0.944 7**	**0.910 5**	**0.915 4**
	MPA	0.846 5	0.872 4	0.924 8	0.590 2	0.812 4	0.612 6
	GNDO	0.894 0	0.906 9	0.944 7	0.637 8	0.815 0	0.687 4
	SMA	0.871 5	0.896 9	0.940 5	0.626 4	0.852 3	0.673 7
	EO	0.874 8	0.901 4	0.952 0	0.753 0	0.864 7	0.741 7
	MRFO	0.868 4	0.892 1	0.943 5	0.757 0	0.810 7	0.714 5

表8 （续）

数据集	算法	fitness	Acc	Precision	F1-score	AUC	Recall
Prostate_Tumor	**BTLBO**	**0.931 9**	**0.955 1**	**0.955 9**	**0.955 4**	**0.958 0**	**0.955 4**
	BTLBOLS	0.912 6	0.950 9	0.952 1	0.951 6	0.952 9	0.951 5
	MPA	0.829 9	0.855 5	0.851 5	0.854 4	0.857 5	0.851 5
	GNDO	0.793 3	0.802 7	0.803 8	0.803 8	0.813 9	0.803 8
	SMA	0.696 6	0.718 2	0.714 2	0.714 9	0.696 3	0.714 2
	EO	0.838 9	0.862 7	0.862 7	0.862 7	0.859 4	0.862 7
	MRFO	0.788 1	0.811 8	0.812 7	0.814 0	0.810 1	0.812 7
SRBCT	**BTLBO**	**0.991 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	BTLBOLS	**0.991 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
	MPA	0.725 0	0.747 2	0.916 0	0.728 5	0.864 3	0.730 4
	GNDO	0.876 8	0.890 3	0.963 0	0.900 4	0.929 5	0.897 3
	SMA	0.594 3	0.612 5	0.865 8	0.624 8	0.738 2	0.572 2
	EO	0.831 4	0.856 9	0.947 9	0.803 8	0.894 6	0.796 1
	MRFO	0.820 6	0.844 4	0.942 4	0.847 3	0.897 7	0.828 1

进化特征选择算法相比都可提高分类精确度. 这也充分说明LS和VNS的高效性,使得算法可以在最优解邻域进行搜索,进一步避免了算法陷入局部最优解.

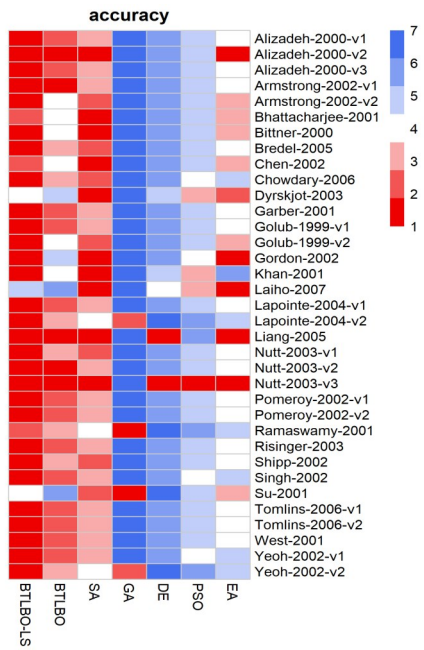

图8 35个基因表达数据集上不同进化算法的精度比较结果

6 总结与展望

近年来,高效的癌症基因在癌症的诊断和治疗中起着至关重要的作用. 然而,由于样本数量有限且远小于数据特征维度,因此从癌症基因表达数据分类中获得这些必需基因是一项艰巨任务. 然而现有方法仍存在一些问题,例如昂贵的计算成本和过早收敛等. 为在全局搜索和局部搜索之间保持良好的平衡,并识别复合疾病的易感基因,我们提出了一种混合进化算法用于基因选择. 首先,本文采用二进制编码方案和综合适应度函数用于寻找疾病相关基因,获得更高的分类精度;然后,在教师阶段为保证种群的优良性,对最差个体实施重启策略;之后,为提高算法性能,避免算法过早收敛,加入局部干扰策略对教与学优化算法的最优解进行干扰. 此外,针对局部搜索干扰失败的情况,引入变邻域搜索机制,直接对整个个体进行干扰,提高种群质量. 为验证算法优越性,实验选取35个癌症基因表达数据集进行测试,并与十种特征选择算法相比. 实验结果表明,所提方法在分类准确度、特征选择率及其他多个指标方面,胜于其他算法. 在今后的工作中,我们将尝试把所提算法运用于肿瘤分析、诊断等真实数据集中,并在实际生物医疗应用中与基于进化计算的特征选择算法进行对比.

参考文献

[1] ZHOU Z H. Machine Learning[M]. Singapore: Springer Singapore, 2021.

[2] HALL M A. Correlation-Based Feature Selection Formachine Learning[D]. Hamilton: The University of Waikato, 1999.

[3] EFRON B, HASTIE T, JOHNSTONE I, et al. Least angle regression[J]. The Annals of Statistics, 2004, 32(2): 407-451.

[4] TIBSHIRANI R, HASTIE T, NARASHMAN B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(10): 6567-6572.

[5] ROBNIK-ŠIKONJA M, KONONENKO I. Theoretical and empirical analysis of ReliefF and RReliefF[J]. Machine Learning, 2003, 53(1): 23-69.

[6] DING C, PENG H C. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2005, 3(2): 185-205.

[7] PENG H C, LONG F H, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and Min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[8] YEOH E J, ROSS M E, SHURTLEFF S A, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling[J]. Cancer Cell, 2002, 1(2): 133-143.

[9] ZHANG J Y, LIU S L, WANG Y. Gene association study with SVM, MLP and cross-validation for the diagnosis of diseases[J]. Progress in Natural Science, 2008, 18(6): 741-750.

[10] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3: 1157-1182.

[11] FALLAHPOUR S, LAKVAN E N, ZADEH M H. Using an ensemble classifier based on sequential floating forward selection for financial distress prediction problem [J]. Journal of Retailing and Consumer Services, 2017, 34: 159-167.

[12] YAP B W, IBRAHIM N, HAMID H A, et al. Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy[J]. Pertanika Journal of Science and Technology, 2018, 26(1): 329-340.

[13] PUDIL P, NOVOVIČOVÁ J, KITTLER J. Floating search methods in feature selection[J]. Pattern Recognition Letters, 1994, 15(11): 1119-1125.

[14] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1): 389-422.

[15] 郭广颂, 陈良骥, 文振华, 等. 求解高维混合指标优化问题的交互式进化计算[J]. 电子学报, 2020, 48(7): 1361-1368.
GUO G S, CHEN L J, WEN Z H, et al. Sloving multidimensional optimization problems with hybird indices by interactive evolutionary computation[J]. Acta Electronica Sinica, 2020, 48(7): 1361-1368. (in Chinese)

[16] 王宇平, 焦永昌, 张福顺. 解无约束非线性全局优化的一种新进化算法及其收敛性[J]. 电子学报, 2002, 30 (12): 1867-1869.
WANG Y P, JIAO Y C, ZHANG F S. A new evolutionary algorithm for unconstrained nonlinear global optimization problems and its convergence[J]. Acta Electronica Sinica, 2002, 30(12): 1867-1869. (in Chinese)

[17] 茅晓泉, 胡光锐, 唐斌. 一种 DHMM 的混合训练方法[J]. 电子学报, 2002, 30(1): 148-150.
MAO X Q, HU G R, TANG B. A hybrid training method for DHMMs [J]. Acta Electronica Sinica, 2002, 30(1): 148-150. (in Chinese)

[18] JIMÉNEZ F, SÁNCHEZ G, GARCÍA J M, et al. Multi-objective evolutionary feature selection for online sales forecasting[J]. Neurocomputing, 2017, 234: 75-92.

[19] JIMÉNEZ F, MARTÍNEZ C, MARZANO E, et al. Multi-objective evolutionary feature selection for fuzzy classification[J]. IEEE Transactions on Fuzzy Systems, 2019, 27 (5): 1085-1099.

[20] MAFARJA M, ALJARAH I, HEIDARI A A, et al. Evolutionary Population Dynamics and Grasshopper Optimization approaches for feature selection problems[J]. Knowledge-Based Systems, 2018, 145: 25-45.

[21] TARADEH M, MAFARJA M, HEIDARI A A, et al. An evolutionary gravitational search-based feature selection [J]. Information Sciences, 2019, 497: 219-239.

[22] NAKISA B, RASTGOO M N, TJONDRONEGORO D, et al. Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors[J]. Expert Systems with Applications, 2018, 93: 143-155.

[23] GHOSH A, DATTA A, GHOSH S. Self-adaptive differential evolution for feature selection in hyperspectral image data[J]. Applied Soft Computing, 2013, 13(4): 1969-1977.

[24] KHUSHABA R N, AL-ANI A, ALSUKKER A, et al. A combined ant colony and differential evolution feature selection algorithm[C]// Proceedings of the 6th International Conference on Ant Colony Optimization and Swarm Intelligence. Berlin: Springer, 2008:1-12.

[25] ZORARPACI E, ÖZEL S A. A hybrid approach of differential evolution and artificial bee colony for feature selection[J]. Expert Systems with Applications, 2016, 62(15): 91-103.

[26] ALLAOUI M, AHIOD B, EL YAFRANI M. A hybrid crow search algorithm for solving the DNA fragment assembly problem[J]. Expert Systems with Applications, 2018, 102: 44-56.

[27] SHUKLA A K, SINGH P, VARDHAN M. Gene selection for cancer types classification using novel hybrid metaheuristics approach[J]. Swarm and Evolutionary Computation, 2020, 54: 100661.

[28] 韩冲, 王俊丽, 吴雨茜, 等. 基于神经进化的深度学习模型研究综述[J]. 电子学报, 2021, 49(2): 372-379.
HAN C, WANG J L, WU Y X, et al. A review of deep learning models based on neuroevolution[J]. Acta Electronica Sinica, 2021, 49(2): 372-379. (in Chinese)

[29] RAO R V, SAVSANI V J, VAKHARIA D P. Teaching-

learning-based optimization: A novel method for constrained mechanical design optimization problems[J]. Computer-Aided Design, 2011, 43(3): 303-315.
[30] WANG Z, LU R Q, CHEN D B, et al. An experience information teaching-learning-based optimization for global optimization[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016, 46(9): 1202-1214.
[31] RAO R V, SAVSANI V J, VAKHARIA D P. Teaching-learning-based optimization: An optimization method for continuous non-linear large scale problems[J]. Information Sciences, 2012, 183(1): 1-15.
[32] ZOU F, WANG L, HEI X H, et al. Teaching-learning-based optimization with dynamic group strategy for global optimization[J]. Information Sciences, 2014, 273: 112-131.
[33] GHASEMI M, GHANBARIAN M M, GHAVIDEL S, et al. Modified teaching learning algorithm and double differential evolution algorithm for optimal reactive power dispatch problem: A comparative study[J]. Information Sciences, 2014, 278: 231-249.
[34] GONZÁLEZ-ÁLVAREZ D L, VEGA-RODRÍGUEZ M A, GÓMEZ-PULIDO J A, et al. Multiobjective teaching-learning-based optimization (MO-TLBO) for motif finding[C]//2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI). New York: Institute of Electrical and Electronics Engineers, 2013: 141-146.
[35] LIN D H, TANG X O. Conditional infomax learning: An integrated framework for feature extraction and fusion [C]//European Conference on Computer Vision. Berlin: Springer, 2006: 68-82.
[36] YANG H H, MOODY J. Data visualization and feature selection: New algorithms for nongaussian data[C]//Proceedings of the 12th International Conference on Neural Information Processing Systems. New York: ACM, 1999: 687-693.
[37] PENG H C, LONG F H, DING C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.
[38] MEYER P E, SCHRETTER C, BONTEMPI G. Information-theoretic feature selection in microarray data using variable complementarity[J]. IEEE Journal of Selected Topics in Signal Processing, 2008, 2(3): 261-274.
[39] ROBNIK- ŠIKONJA M, KONONENKO I. Theoretical and empirical analysis of ReliefF and RReliefF[J]. Machine Learning, 2003, 53(1): 23-69.
[40] FARAMARZI A, HEIDARINEJAD M, MIRJALILI S, et al. Marine predators algorithm: A nature-inspired metaheuristic[J]. Expert Systems with Applications, 2020, 152: 113377.
[41] ZHANG Y Y, JIN Z G, MIRJALILI S. Generalized normal distribution optimization and its applications in parameter extraction of photovoltaic models[J]. Energy Conversion and Management, 2020, 224: 113301.
[42] LI S M, CHEN H L, WANG M J, et al. Slime mould algorithm: A new method for stochastic optimization[J]. Future Generation Computer Systems, 2020, 111: 300-323.
[43] FARAMARZI A, HEIDARINEJAD M, STEPHENS B, et al. Equilibrium optimizer: A novel optimization algorithm[J]. Knowledge-Based Systems, 2020, 191: 105190.
[44] ZHAO W G, ZHANG Z X, WANG L Y. Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications[J]. Engineering Applications of Artificial Intelligence, 2020, 87: 103300.
[45] ZHAO W G, ZHANG Z X, WANG L Y. Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications[J]. Engineering Applications of Artificial Intelligence, 2020, 87: 103300.

作者简介

高慧敏 女，1997 年 3 月生，山西大同人．吉林大学人工智能学院硕士研究生．主要研究方向为进化计算和特征选择．

王云鹤 女，1991 年 4 月生，河北沧州人．现为河北工业大学人工智能与数据科学学院讲师．主要研究方向为智能计算和机器学习．

卞 闯 男，1996 年 7 月生，吉林长春人．吉林大学人工智能学院硕士研究生．主要研究方向为生物信息学和特征选择．

李向涛（通讯作者） 男，1987 年 4 月生，江苏淮安人．现为吉林大学人工智能学院教授．主要研究方向为智能计算、进化数据挖掘、约束优化、多目标优化及其应用．
E-mail: lixt314@jlu.edu.cn