



# 改进的基于层次距离的基因表达式编程特征选择分类算法

湛 航<sup>1</sup>, 何 朗<sup>1\*</sup>, 黄樟灿<sup>1</sup>, 李华峰<sup>1</sup>, 张 蓄<sup>1</sup>, 谈 庆<sup>2</sup>

(1. 武汉理工大学 理学院, 武汉 430070; 2. 武汉大学 数学与统计学院, 武汉 430072)

(\* 通信作者电子邮箱 helang@whut.edu.cn)

**摘 要:**针对一般特征选择算法未能揭示数据特征与数据类别之间的可解释性映射关系的问题,在基因表达式编程(GEP)的基础上,通过引入初始化方法、变异策略以及适应度评价方法,提出了一种改进的基于层次距离的GEP特征选择分类算法(FSLDGE)。首先,利用定义的选择概率有导向地初始化种群个体,从而增加种群中有效个体的数量;其次,定义个体的层次邻域,使种群个体基于其层次邻域进行变异,并解决了变异过程中的盲目无导向性问题;最后,将维度缩减率与分类准确率结合起来作为个体的适应度值,从而改变种群单一优化目标的进化模式,并平衡两者之间的关系。在7个数据集上进行5折交叉和10折交叉验证,所提算法给出了数据特征及其类别之间的函数映射关系,将得到的映射函数用于数据分类。与森林优化特征选择算法(FSFOA)、邻域软边界特征选择算法(NSM)、基于邻域有效信息比的特征选择算法(FS-NEIR)等对比算法相比,所提算法的维度缩减率在Hepatitis、WPBC(Wisconsin Prognostic Breast Cancer)、Sonar、WDBC(Wisconsin Diagnostic Breast Cancer)数据集上得到了最好结果;与对比算法相比,所提算法的平均分类准确率在Hepatitis、Ionosphere、Musk1、WPBC、Heart-Statlog、WDBC数据集上得到了最好结果。实验结果验证了所提算法在特征选择分类问题上的可行性、有效性和优越性。

**关键词:**特征选择;函数发现;基因表达式编程;种群初始化;层次邻域

**中图分类号:**TP317.4 **文献标志码:**A

## Improved feature selection and classification algorithm for gene expression programming based on layer distance

ZHAN Hang<sup>1</sup>, HE Lang<sup>1\*</sup>, HUANG Zhangcan<sup>1</sup>, LI Huafeng<sup>1</sup>, ZHANG Qiang<sup>1</sup>, TAN Qing<sup>2</sup>

(1. School of Science, Wuhan University of Technology, Wuhan Hubei 430070, China;

2. School of Mathematics and Statistics, Wuhan University, Wuhan Hubei 430072, China)

**Abstract:** Concerning the problem that the interpretable mapping relationship between data features and data categories do not be revealed by general feature selection algorithms. on the basis of Gene Expression Programming (GEP), by introducing the initialization methods, mutation strategies and fitness evaluation methods, an improved Feature Selection classification algorithm based on Layer Distance for GEP (FSLDGE) was proposed. Firstly, the selection probability was defined to initialize the individuals in the population directionally, so as to increase the number of effective individuals in the population. Secondly, the layer neighborhood of the individual was proposed, so that each individual in the population would mutate based on its layer neighborhood, and the blind and unguided problem in the process of mutation was solved. Finally, the dimension reduction rate and classification accuracy were combined as the fitness value of the individual, which changed the population evolutionary mode of single optimization goal and balanced the relationship between the above two. The 5-fold and 10-fold verifications were performed on 7 datasets, the functional mapping relationship between data features and their categories was given by the proposed algorithm, and the obtained mapping function was used for data classification. Compared with Feature Selection based on Forest Optimization Algorithm (FSFOA), feature evaluation and selection based on Neighborhood Soft Margin (NSM), Feature Selection based on Neighborhood Effective Information Ratio (FS-NEIR) and other comparison algorithms, the proposed algorithm has obtained the best results of the dimension reduction rate on Hepatitis, Wisconsin Prognostic Breast Cancer (WPBC), Sonar and Wisconsin Diagnostic Breast Cancer (WDBC) datasets, and has the best average classification accuracy on Hepatitis, Ionosphere, Musk1, WPBC, Heart-Statlog and WDBC datasets. Experimental results shows that the feasibility, effectiveness and superiority of the proposed algorithm in feature selection and classification are verified.

**Key words:** feature selection; function discovery; Gene Expression Programming (GEP); population initialization; layer neighborhood

收稿日期: 2020-11-17; 修回日期: 2021-03-09; 录用日期: 2021-03-10。 基金项目: 国家自然科学基金面上项目(61672391)。

作者简介: 湛航(1996—), 男, 湖北孝感人, 硕士研究生, 主要研究方向: 智能计算; 何朗(1974—), 男, 湖北鄂州人, 教授, 博士, 主要研究方向: 演化计算、智能计算; 黄樟灿(1960—), 男, 浙江嵊州人, 教授, 博士, 主要研究方向: 智能计算、图像处理; 李华峰(1995—), 男, 山西临汾人, 硕士研究生, 主要研究方向: 智能计算; 张蓄(1996—), 女, 河南信阳人, 硕士研究生, 主要研究方向: 智能计算; 谈庆(1991—), 男, 湖南长沙人, 博士研究生, 主要研究方向: 智能计算、图像处理。



## 0 引言

特征选择 (Feature Selection, FS) 是数据挖掘中一项特别重要的数据预处理步骤, 常用于去除数据中冗余和不相关的特征数据, 降低计算复杂度, 提升模型的效果<sup>[1-2]</sup>。特征选择在数字图像处理、文本分类、生物信息学、基因序列分析、金融时间序列分析以及医学数据分析等方面应用较为广泛<sup>[3-6]</sup>。常见的特征选择方法有: 过滤式 (filter)、嵌入式 (embedding) 和包裹式 (wrapper)<sup>[7]</sup>。过滤式独立于学习器, 在训练学习器之前完成特征的选择; 嵌入式则将特征选择与学习器的训练结合, 在一个优化过程中完成; 包裹式则依据学习器的性能作为特征选择优劣的评价<sup>[7]</sup>。

特征选择是一个 NP-hard 问题, 对于具有  $n$  个特征的特征选择问题有  $2^n$  个可能的特征子集<sup>[8]</sup>。在分类问题上, 通过选取关键的特征, 忽略不太重要的数据特征, 构建出一个良好的分类器, 进而提高分类预测效果<sup>[9]</sup>。选择的特征优劣对于构建出一个简洁、高效的分类系统有着重要的影响<sup>[10]</sup>。

国内外许多学者一直致力于研究这个问题。一些简单方法如穷举搜索算法、分支定界算法等<sup>[11]</sup>, 这些方法可以挑选出最优的特征子集, 但是这些方法的时间复杂度对于高维数据并不友好, 在应用过程中也存在一些局限性。基于该问题, 一些传统的机器学习方法被应用到这个方面, 如决策树算法<sup>[12]</sup>给出了一种基于信息增益大小的特征选择分类方法; 李占山等<sup>[7]</sup>提出了一种基于 XGBoost 的特征选择算法; Moustakidis 等<sup>[13]</sup>提出了一种基于支持向量机 (Support Vector Machine, SVM) 的模糊互补准则的特征选取方法。虽然机器学习方法拥有较快的求解速度, 模型容易解释, 但是它没有考虑特征与特征之间、特征与类别之间相关性。Sun 等<sup>[14]</sup>结合类别相关度, 将最大相关性和最小冗余准则 (max-Relevance and min-Redundancy, mRMR) 扩展到了多标签学习, 借助凸优化来优化特征初始分配的权重, 从而获取特征排序; Sheikhi 等<sup>[15]</sup>将特征的冗余性替换为多样性, 以量化候选特征相较于已选择子集的互补性, 提出了正秩最大相关性及多样性的特征选择算法。尽管这些方法的原理较为简单, 在小规模数据上拥有不错的表现, 但对于大规模数据而言, 在求解速度与求解精度上略显不足。

由于演化算法在 NP-hard 问题上拥有不错的表现, 其计算复杂度相较于其他机器学习等方法更低, 对于整个问题的解给出一种编码方案, 而不是直接对问题的具体参数进行处理, 它不用考虑问题本身的特殊知识背景, 仅依靠种群个体内部的进化就能寻找到一个全局的较优解<sup>[7]</sup>, 因而被大量用于特征选择问题上。Ghaemi 等<sup>[16]</sup>将求解连续变量优化问题的森林优化算法应用于离散的特征选择问题, 有效地提高了分类器的分类精度; 张翠军等<sup>[1]</sup>提出了一种平衡特征子集个数与分类精度策略的多目标骨架粒子群优化的特征选取算法; 张梦林等<sup>[17]</sup>提出一种采用新的初始化策略和评估函数, 以特征子集的准确率指导采样的基于离散域采样分类优化 (Sampling-And-Classification optimization, SAC) 特征选择算法, 实验结果表明该方法可提高分类准确率, 有好的泛化能力; 李光华等<sup>[18]</sup>将随机森林的重要度评分作为蚁群的启发式信息, 提出了一种融合蚁群算法和随机森林的特征选择算法; Tabakhi 等<sup>[19]</sup>提出一种不依赖于其他学习算法, 通过多次迭代寻优寻找最优子集的基于蚁群优化的无监督特征选择算法。

演化算法在特征选择领域的应用, 使得人们不再关注数

据领域内的具体知识, 通过控制演化算法中种群的规模和搜索策略, 采取合适的分类器, 使得演化算法取得了不错的效果<sup>[11]</sup>。但是, 上述演化算法往往还需要结合其他分类器才能在特征选择问题上使用, 演化算法仅仅只是选择了特征, 当前特征选择的优劣与否还需要借助分类器分类效果进行判断。并且, 选择的特征与分类器仅仅只能反映出当前特征与该数据所属类别之间存在关系, 无法直观说明关键特征与数据类别之间的一种可解释的映射关系。

基于该问题, Ma 等<sup>[20-21]</sup>提出了一种基于遗传规划 (Genetic Programming, GP) 算法的特征选择算法, 利用 GP 算法在函数发现上的优越性, 构建出特征与类别之间的函数关系。实验结果表明, 该算法提高了分类准确度, 降低了特征选择率, 还给出了显式函数关系式, 揭示了特征与类别之间存在的函数关系。但是 GP 算法存在膨胀现象, 种群个体在进化的过程中容易出现无用个体, 影响种群进化效率, 而基因表达式编程 (Gene Expression Programming, GEP) 算法相较于 GP 算法拥有更加高效的函数发现能力, 能更好地克服 GP 中的缺陷<sup>[22]</sup>。GEP 算法在演化过程中过多的遗传算子参数设定使得算法结果容易受到干扰, 而且决定演化方向的单一适应度值会使算法在解决复杂问题时容易陷入局部最优。崔未<sup>[23]</sup>提出了基于层次距离的基因表达式编程 (GEP based on Layer Distance, LDGEP) 算法, 通过改变染色体结构, 对个体的基因进行分层, 定义种群个体层与层之间的距离, 通过层次距离来选择遗传位点, 使得个体有针对性地变异。实验结果表明, LDGEP 算法相较于传统 GEP 算法有更好的发现函数的能力与速度, 但在种群初始化及层次变异过程中存在一定的盲目性, 影响种群进化速度。

基于此, 本文提出了改进的基于层次距离的基因表达式编程特征选择分类算法 (improved Feature Selection and classification algorithm for LDGEP, FSLDGEP), 通过改进种群的初始化方式, 使得种群个体的随机初始化具有一定的导向性; 同时, 通过定义种群的层次邻域来改变种群个体的随机变异方式; 并在原始的分类准确率评价指标上, 考虑种群个体选择的特征数量, 将其作为适应度评价指标之一, 平衡维度缩减率与分类准确率之间的关系, 构建出类别关于特征的函数, 揭示出两者之间存在的函数映射关系。

## 1 基于层次距离的基因表达式编程算法

LDGEP 算法是由崔未基于 GEP 算法改进而来的群智能演化算法。传统的 GEP 算法中存在许多需要手动设定参数的遗传进化算子, 如选择复制算子、转座算子、重组算子和变异算子等, 而 LDGEP 算法更改了原始 GEP 算法的演化模式, 重新定义了个体结构与个体相似性计算方法, 将大量的进化算子减少至三个算子, 且算子的参数无需设置, 提升了算法求解问题的精度和速度。LDGEP 算法主要包括四步: 种群初始化、适应度计算、选择操作和遗传操作。

### 1.1 种群初始化与个体编码解码

LDGEP 算法中初始种群个体的基因是随机生成的, 基因由一个头部和尾部组成, 其长度分别为  $h$  和  $t$ 。头部符号从函数集和终止集中选取, 尾部符号从终止集中选取。一般的函数集包含常用的数学运算符及一些初等函数, 如  $\{+, -, *, /, \sin, \exp, \dots\}$ , 终止集则通常由变量、常数等组成, 如  $\{x_1, x_2, 1, 2, e, \dots\}$ 。当基因头部与基因尾部长度的满足约束式 (1) 时, 个体初始化完成, 重复多次, 即可完成种群的初

始化。

$$t = h * (n - 1) + 1 \quad (1)$$

其中: $t$ 表示基因尾部长度; $h$ 表示基因头部长度; $n$ 表示函数集中符号的最大运算目数。

传统 GEP 算法中的个体采用固定长度编码,只是在表达时部分基因不表达。为确保个体间的距离定义一致性以及表达式树结构的完整性,LDGEP 算法对未表达的节点采用 TR (Null)节点补齐方法。LDGEP 算法中的个体编码长度由该个体的层数确定,其长度计算公式如式(2):

$$len = 2^L - 1 \quad (2)$$

其中: $len$ 表示个体长度; $L$ 表示层数。

如表达式  $a + \sin(b)$ ,该个体的编码及其表达式树结构如图 1(a)。解码时,根据运算目数对未表达的基因采用 TR 补齐,如图 1(b),虚线代表补齐的 TR 节点。再根据各节点运算符的性质以及二叉树的中序遍历方式,将个体的树结构转换为数学表达式。

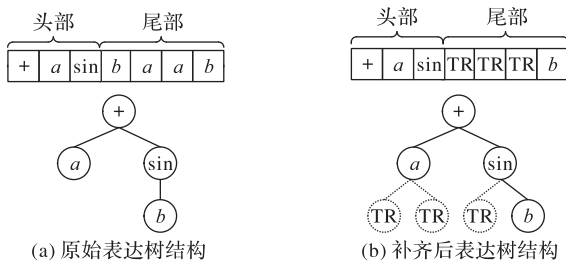


图 1 表达式树结构及补齐后表达式树结构

Fig. 1 Structures of expression tree and complemented expression tree

## 1.2 距离定义

种群个体的层次示意图如图 2 所示,每层个体的数量为  $2^{l-1}$ ,其中  $l$  表示当前的层数。

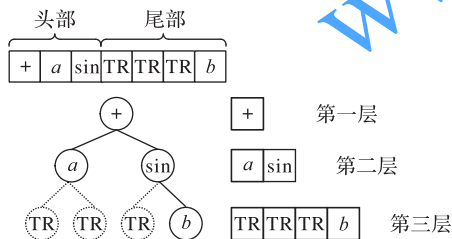


图 2 个体层次示意图

Fig. 2 Schematic diagram of individual layer

LDGEP 算法对每个个体的每一层给定一个权值,用以区分不同层对于个体的影响。种群个体的层次距离为基于加权的汉明距离,计算公式所示:

$$d_{ij}(l) = w_l \cdot D_{ij}(l); l = 1, 2, \dots, L \quad (3)$$

其中: $d_{ij}(l)$ 表示第  $i$  个个体与第  $j$  个个体在第  $l$  层的层次距离; $L$ 表示个体的总层数; $D_{ij}(l)$ 为第  $i$  个个体与第  $j$  个个体在第  $l$  层的汉明距离。 $w_l$ 表示第  $l$  层的权重,其计算公式如下所示:

$$w_l = 1/2^{l-1} \quad (4)$$

其中: $l$ 表示层数。随着层次的递增,层次重要性越低,从而对解码后的个体的表达影响越低。

## 1.3 遗传操作

### 1.3.1 遗传位点选择

由于每一层对于种群个体的表达有不同的影响,通过对层次距离不同的层采用遗传操作,使种群个体向更优秀的个

体靠近。通过计算每一层个体间的层次距离确定遗传位点。遗传位点的确定规则如下所示:

$$l = \begin{cases} l + 1, & d_{ij}(l) < 0.5 \\ layer_{ij}, & d_{ij}(l) \geq 0.5 \end{cases} \quad (5)$$

其中: $d_{ij}(l)$ 表示第  $i$  个个体与第  $j$  个个体在第  $l$  层的层次距离; $layer_{ij}$ 表示第  $i$  个个体与第  $j$  个个体遗传位点层。若该层层次距离小于 0.5,则忽略该层,转入下一层。否则,用  $layer_{ij}$ 记录下当前所处层,并对其执行三种遗传操作:1)对两个个体的第  $l+1$  层及之后剩余层进行基因重组;2)对两个个体的第  $l$  层进行层间重组;3)对该个体的第  $l+1$  层及之后剩余层进行层次变异。

### 1.3.2 基因重组

根据式(5),确定遗传选择层,其操作机制如图 3 所示,两个个体第一层的层次距离大于 0.5,确定遗传位点为第二层,基因重组算子对两个个体剩余部分执行重组。

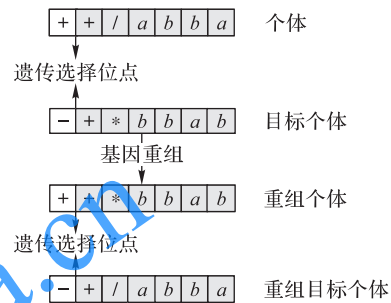


图 3 基因重组

Fig. 3 Gene recombination

### 1.3.3 层间重组

根据式(5),确定遗传选择层,层间算子对该层执行层间重组。其操作机制如图 4,两个个体的第一层层次距离为 0,第二层层次距离等于 0.5,对两个个体的第二层进行重组。

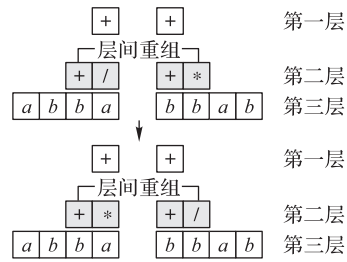


图 4 层间重组

Fig. 4 Recombination between layers

### 1.3.4 层次变异

为避免无用个体的产生,执行层次变异时,第一层符号从函数集选择,最后一层符号从终止集选择,而其他层则从函数集与终止集中选择。根据式(5),确定遗传选择层,其操作机制如图 5 所示,第一层距离大于 0.5,层次变异算子从第二层开始执行变异,于是该个体随机从函数集和终止集中选择了一个符号,使得第二层的一个基因由双目运算符变成单目运算符,同时为保证个体正确表达,对其子节点进行 TR 补齐。

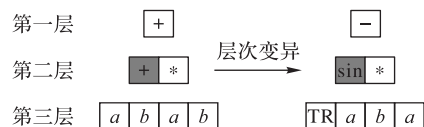


图 5 层次变异

Fig. 5 Layer mutation





## 2 改进的LDGEP特征选择分类算法

由于LDGEP算法的种群个体有着特殊的基因编码、解码方式,因此,当个体的头部基因有较多终止符时,容易产生一些无用个体。同时,种群个体随机地从函数集和终止集中选择符号来执行层次变异操作,这种随机性减缓了种群的进化速度。针对上述LDGEP算法的不足,本文提出了一种依概率种群初始化方式和基于邻域的层次变异策略,同时,为减少特征的选择数量,将个体选择的特征维度缩减率作为判断最优个体的指标之一。该算法流程如图6所示。

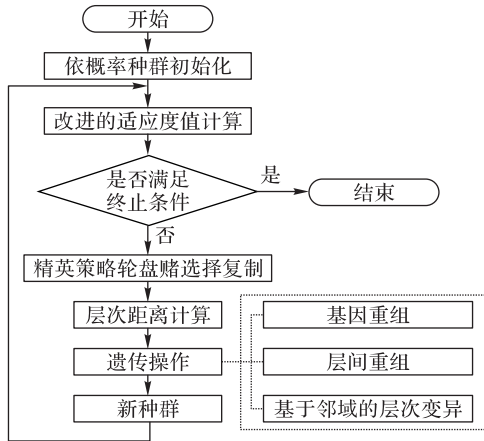


图6 FSLDGE流程

Fig. 6 Flowchart of FSLDGE algorithm

### 2.1 依概率种群个体初始化

在LDGEP算法中,每层符号的选择随着层次的不同而不同。但是,个体编码中过多的终止集符号使得个体的表达提前结束,容易产生无用的表达式。本文提出了依概率种群个体初始化方法,通过概率约束,使个体基因头部更侧重于选择函数符号,增加产生有效表达式个体的数量,提高种群的进化速度。选择概率的计算方式如式(6):

$$P_{select} = \frac{\text{length}(End\_symble)}{\text{length}(Symble)} \quad (6)$$

其中: $P_{select}$ 表示选择概率; $End\_symble$ 表示终止集; $Symble$ 表示函数集和终止集的合集; $\text{length}(\cdot)$ 表示取集合的长度。

依概率种群个体初始化方法如下所示:

输入 种群大小  $N$ , 个体基因编码长度  $len$ , 个体的层数为  $L$ , 基因选择概率  $P_{select}$ , 函数集  $Func\_symble$ , 终止集  $End\_symble$ , 函数集与终止集的合集  $Symble$ 。

输出 已初始化种群  $pop$ 。

- 1) FOR  $i=1$  TO  $N$  DO
- 2) FOR  $j=1$  TO  $len$  DO
- 3) 生成随机数  $r=\text{rand}()$
- 4) 计算当前基因层次  $l=\text{floor}(\log_2(j))$
- 5) IF  $r > P_{select}$  AND  $l < L$  DO
- 6) 被选择的符号在符号集中的位置  $PossibleIndex = \text{round}(\text{rand} * \text{length}([Symble]))$
- 7) 选择当前位置符号作为个体基因编码符号  $pop(i,j) = Symble(PossibleIndex)$
- 8) ELSE IF  $l < L$  DO
- 9) 被选择的符号在符号集中的位置  $PossibleIndex = \text{round}(\text{rand} * \text{length}([Func\_symble]))$
- 10) 选择当前位置符号作为个体基因编码符号  $pop(i,j) = Func\_symble(PossibleIndex)$

- 11) ELSE DO
- 12) 被选择的符号在符号集中的位置  $PossibleIndex = \text{round}(\text{rand} * \text{length}([End\_symble]))$
- 13) 选择当前位置符号作为个体的基因编码符号  $pop(i,j) = End\_symble(PossibleIndex)$
- 14) ENDIF
- 15) ENDFOR
- 16) ENDFOR

### 2.2 基于邻域的层次变异

LDGEP算法针对不同层产生的层次变异,其变异选择符号集不同。而变异的随机性使得种群个体在进化过程中容易产生低质量的解,影响算法的寻优速度。本文提出了层次邻域的概念,层次邻域是由种群中优秀个体的各层次符号组成,当种群个体发生层次变异时,个体基因从当前层次邻域中选取变异符号,使变异个体的基因向着优秀个体基因的方向靠近,提升算法的寻优能力。层次邻域定义如下。

设种群中的个体数为  $N$ , 层数为  $L$ , 计算个体的适应度值, 并从大到小排列, 取前  $N/2$  个个体作为优秀个体, 设前  $N/2$  个个体及其基因编码为:

$$\begin{cases} x_1(g_{1,1}, g_{1,2}, \dots, g_{1,2^L-1}) \\ x_2(g_{2,1}, g_{2,2}, \dots, g_{2,2^L-1}) \\ \vdots \\ x_{N/2}(g_{N/2,1}, g_{N/2,2}, \dots, g_{N/2,2^L-1}) \end{cases} \quad (7)$$

将个体  $x_1, x_2, \dots, x_{N/2}$  中每一层的符号提取去重后构成的集合作为该层的一个层次邻域, 因此可以得到  $L$  个层次邻域式(8):

$$\begin{cases} layer_1: \{g'_{1,1}, g'_{1,2}, \dots, g'_{1,m_1}\} \\ layer_2: \{g'_{2,1}, g'_{2,2}, \dots, g'_{2,m_2}\} \\ \vdots \\ layer_L: \{g'_{L,1}, g'_{L,2}, \dots, g'_{L,m_L}\} \end{cases} \quad (8)$$

其中:  $m_1, m_2, \dots, m_L$  分别表示去重后每一层的符号总个数, 即该层次邻域的大小。种群个体从每一层的层次邻域中选择符号进行层次变异。基于层次邻域的变异策略如下所示:

输入 变异选择位点  $mutation$ , 种群  $pop$ , 层次邻域  $Layer$ , 种群大小为  $N$ , 个体层数  $L$ 。

输出 层次变异后的种群  $newPop$ 。

- 1) FOR  $i=1$  TO  $N$  DO
- 2) 起始变异层次  $l=mutation(i)$ ;
- 3) FOR  $j=2^l$  TO  $2^L-1$  DO
- 4) 符号所在层次  $layerIndex = \text{floor}(\log_2(j))+1$
- 5) 层次变异选择符号索引  $genIndex = \text{floor}(\text{rand} * \text{length}(Layer(layerIndex)))$
- 6) 改变当前符号  $pop(i,j) = Layer(layerIndex, genIndex)$
- 7) ENDFOR
- 8) ENDFOR
- 9)  $newPop = pop$

### 2.3 改进的适应度值计算

LDGEP算法仅将分类准确率  $CA$  (Classification Accuracy) 作为判断当前个体是否最优的指标, 而忽视特征维度缩减问题, 导致算法得到的函数表达式变得复杂的同时也让表达式对特征数据较为依赖和敏感。为了提高算法对于数据特征选择的性能, 本文引入特征的维度缩减率  $DR$  (Dimension Reduction) 作为新的适应度评价指标之一, 将其与  $CA$  加权求



和后的值作为优化目标,改变种群单一优化目标的局限性,平衡了分类函数分类准确率和维度缩减率。适应度值计算方式如下:

$$fitness = \lambda_1 CA + \lambda_2 DR \quad (9)$$

其中: $fitness$ 表示个体的适应度值; $\lambda_1$ 表示准确率的权重; $\lambda_2$ 表示维度缩减率的权重。为保证分类准确率在适应度值中的主导地位,避免维度缩减率的影响, $\lambda_1$ 与 $\lambda_2$ 满足 $\lambda_1 \gg \lambda_2 > 0$ 。

$CA$ 描述的是分类器分类效果的好坏,计算公式如下:

$$CA = (TP + TN) / (TP + TN + FP + FN) \quad (10)$$

其中: $TP$ (True Positives)表示被模型分类为正的正样本数; $TN$ (True Negatives)表示被模型分类为负的负样本数; $FP$ (False Positives)表示被模型分类为负的正样本数; $FN$ (False Negatives)表示被模型分类为正的负样本数。

特征选择数量与 $DR$ 之间满足关系式(11):

$$DR = 1 - SelectFeature / AllFeature \quad (11)$$

其中: $SelectFeature$ 表示个体选择的特征个数; $AllFeature$ 表示数据集的总特征数。故 $DR$ 可替代特征选择数量来描述算法在特征选择上的优劣。

$CA$ 描述的是通过选择的特征构造的分类函数对数据分类正确的概率; $DR$ 描述的是维度缩减率,即未选择的特征在总体特征中的占比。特征选择的数量越小越好,而 $CA$ 越大越好,两者共同用来描述算法性能优劣时评价不统一,又 $DR$ 由选择特征的数量决定,且 $DR$ 的大小变化与描述算法性能优劣一致,故采用 $DR$ 替代特征选择数量来侧面描述算法在特征选择上的优劣。 $CA$ 与 $DR$ 共同构成本文算法适应度值计算的相关指标。

## 2.4 算法步骤

第1步 初始化参数。确定种群大小 $NIND$ ,种群迭代次数 $MAXGEN$ ,终止集 $END\_SYMBLE$ ,函数集 $FUNC\_SYMBLE$ ,个体编码层数 $L$ 。

第2步 依概率种群初始化。根据式(6)计算选择概率,根据依概率种群初始化算子从 $FUNC\_SYMBLE$ 和 $END\_SYMBLE$ 中选择 $2^l$ 个符号作为个体 $X$ ,构造初始种群。

第3步 种群适应度计算。将种群个体的编码转换成对应的分类函数,通过计算函数中特征数据变量的个数以及利用分类函数对数据进行分类,得到维度缩减率及分类准确率,再根据式(9)计算每个个体的适应度值。

第4步 个体更新。对个体执行精英策略轮盘选择复制算子、基因重组、层间重组算子以及基于邻域的层次变异等遗传操作,得到更新后的个体 $NewX$ 。

第5步 终止条件判断。如果 $gen > MAXGEN$ ,则终止,得到分类函数、维度缩减率以及分类准确率;否则令 $gen = gen + 1, X = NewX$ ,返回第3步。

## 2.5 时间复杂度分析

设最大迭代次数为 $MAXGEN$ ,种群规模为 $NIND$ ,种群个体编码长度为 $l$ ,数据样本量 $M$ ,由2.4节算法步骤可知,在一个完整的循环过程中:第1步与第5步的时间复杂度可以忽略不计;第2步中由于需要遍历个体 $X$ ,针对长度为 $l$ 的个体 $X$ ,初始化过程的时间复杂度近似为 $O(NIND \cdot l)$ ;第3步操作由于每个个体均需要计算一次样本数据的结果,其时间复杂度近似为 $O(NIND \cdot M)$ ;第4步中均是个体自身以及个体与个体之间的遗传操作,其最差情况下的时间复杂度为 $O(NIND \cdot l)$ 。因此本文算法的总体时间复杂度近似可表示为 $O(MAXGEN \cdot NIND \cdot (M + l) + NIND \cdot l)$ 。

## 3 仿真实验

### 3.1 实验数据

实验采用UCI<sup>[24]</sup>机器学习库中的7个分类数据集Hepatitis、Ionosphere、Musk1、Sonar、Heart-Statlog、WDBC (Wisconsin Diagnostic Breast Cancer)和WPBC (Wisconsin Prognostic Breast Cancer)。其中Hepatitis数据中存在多个数据缺失的情况,本文利用其同类特征数据的平均值替代。为避免不同特征数据的量纲影响,本文对7组数据分别进行了归一化处理,归一化方法如式(12)所示:

$$F'_{value} = \frac{F_{value} - F_{min}}{F_{max} - F_{min}} \quad (12)$$

其中: $F'_{value}$ 表示归一化后特征数据值; $F_{value}$ 表示未归一化特征数据值; $F_{min}$ 表示未归一化特征数据中的最小值; $F_{max}$ 表示未归一化特征数据中的最大值。

数据集详细信息如表1描述。实验仿真在Intel Core i5-4460 3.20 GHz CPU, 4.0 GB 内存, 64 位 Windows 7 操作系统的PC上实现,软件环境为Matlab R2017b。

表1 实验数据集说明

Tab. 1 Description of experimental datasets

| 数据名称          | 数据个数 | 特征数 | 类别 | 说明    |
|---------------|------|-----|----|-------|
| Hepatitis     | 155  | 19  | 2  | 肝炎    |
| Ionosphere    | 351  | 34  | 2  | 电离层雷达 |
| Musk1         | 476  | 166 | 2  | 麝香分子  |
| WPBC          | 194  | 33  | 2  | 乳腺癌病例 |
| Sonar         | 208  | 60  | 2  | 声呐信号源 |
| Heart-Statlog | 270  | 13  | 2  | 心脏病   |
| WDBC          | 569  | 30  | 2  | 乳腺癌病例 |

### 3.2 实验参数

FSLDGEF参数设置为:种群大小40,种群迭代次数300,个体层数的值如表2,常数个数10以及函数集 $\{+, -, *, /, \text{sqrt}, \text{exp}, \text{sin}, \text{cos}, \text{tan}, \text{lg}\}$ ,终止集由各类特征数据和常数组成。经过多组数据实验结果比较,将式(9)中 $\lambda_1$ 的权重值设定为1, $\lambda_2$ 权重值设定为0.01。

表2 实验数据集参数

Tab. 2 Parameters of experimental datasets

| 数据名称          | 特征个数 | 个体层数 |
|---------------|------|------|
| Hepatitis     | 19   | 7    |
| Ionosphere    | 34   | 8    |
| Musk1         | 166  | 10   |
| WPBC          | 33   | 8    |
| Sonar         | 60   | 7    |
| Heart-Statlog | 13   | 6    |
| WDBC          | 30   | 6    |

为了防止产生过拟合的问题,本文采用2折交叉、5折交叉、10折交叉与70%-30%这4种验证方式。其中, $k$ 折交叉验证方式( $k$ 表示2、5、10)表示:将数据集分成 $k$ 份,训练时每次取 $k-1$ 份数据做训练,剩余1份数据做测试,使得每一份数据都能基于 $k-1$ 份数据训练得到的结果做测试,验证算法性能。70%-30%验证方式表示取原始数据集中的70%的数据作为训练集,训练算法;30%的数据作为测试集,验证算法性能。

对于分类问题,本文采用IF-THEN规则<sup>[23]</sup>,根据算法输出



结果与阈值比较,判断出当前数据特征所属类别,IF-THEN 规则如下:

if  $outValue > threshold$  then  $class = 1$ , else  $class = 0$ .

其中: $outValue$ 表示输出结果; $threshold$ 表示阈值,在本文中,阈值的大小设定为0; $class$ 表示数据类别。

评价算法的性能主要使用两个指标<sup>[17]</sup>:  $DR$ 和 $CA$ 。 $CA$ 越大,算法得到的映射函数的分类性能越好; $DR$ 越大,则算法选择特征数量越少,维度缩减能力越强,特征选择能力越强。

### 3.3 实验结果

为验证本文算法对于构建特征及其类别之间函数映射关

系的可行性,本文对不同数据在不同的验证方式下独立重复了10次实验,取10次实验中测试集的最优分类准确率 $TBCA$ (Testset Best CA)和最优分类准确率对应的维度缩减率 $TBDR$ (Testset Best DR)及其相对应的分类函数 $TBF$ (Testset Best Function),其结果展示如表3所示。

表3的最优分类函数 $TBF$ 及式(13)中的 $F_i$ 分别表示数据中的第 $i$ ( $i = 1, 2, \dots, M$ ,  $M$ 表示数据的维度)个特征数据,该特征数据与原始数据集中的特征数据一一对应。 $c_7$ 表示常数。将对应的特征数据带入表5中的最佳分类函数中进行计算,即可判断出当前特征数据所属类别。

表3 最优分类准确率及维度缩减率对应函数

Tab. 3 Functions corresponding to optimal classification accuracy and dimension reduction rate

| 数据集           | 验证方式    | $TBCA/\%$ | $TBDR/\%$ | $TBF$  |
|---------------|---------|-----------|-----------|--|
| Hepatitis     | 10-fold | 100.00    | 89.47     | $\exp(\sin^2(F_{17}) - \exp(\sin(F_{12}))) - \exp(2F_{17})$                            |
| Hepatitis     | 70%-30% | 91.30     | 89.47     | $\sin(\sin(F_{12}^2)/(F_{12} - F_{14}))$   |
| Ionosphere    | 10-fold | 100.00    | 85.29     | $\sin(\sin(F_{27} + F_{30})^2)\cos(F_2 + F_5 + F_8 + \sin(\exp(\sin(\tan(F_{27}))))))$ |
| Ionosphere    | 70%-30% | 94.29     | 91.18     | $F_4^2 - \sin(\tan(\sin(F_3 F_5)))$  |
| Ionosphere    | 2-fold  | 90.34     | 88.24     | $F_5^2 F_7^2 \tan(\sin(F_{22} - F_5))\cos(F_{32})$                                     |
| Musk1         | 10-fold | 83.33     | 86.75     | 式(13)  |
| WPBC          | 5-fold  | 87.50     | 84.85     | $\exp(F_3 \tan(\exp(F_8^2))) - \tan(\sin(F_{15})) + \exp(F_4) \tan(F_1)$               |
| Sonar         | 10-fold | 95.24     | 95.00     | $\lg(F_{42} F_{50}) - \tan(F_{12})$  |
| Sonar         | 70%-30% | 82.26     | 95.00     | $\cos\left(\sqrt[4]{F_{52}/c_7}\right)(\lg(F_{11}) + \sqrt{F_{13}}), c_7 = 2.25$       |
| Sonar         | 2-fold  | 75.00     | 93.33     | $\cos(\exp(\cos(F_{52}) - \sin(\sqrt{F_9}) + \cos(F_{48}))) + \sqrt{\exp(F_{35})})$    |
| Heart-Statlog | 10-fold | 92.59     | 76.92     | $\cos(\sin(F_{13} + \cos(F_{12}))) + \exp(\sin(F_3 + F_{12}))$                         |
| Heart-Statlog | 2-fold  | 82.96     | 69.23     | $\tan(F_8) - F_3 - F_{12} \exp(F_{12}/F_{13})$   |
| WDBC          | 10-fold | 100.00    | 86.67     | $\sin(F_2 + F_{21} + F_{28} + \cos(F_{21})) - \sin(F_{21}) \tan(F_{29})$               |
| WDBC          | 5-fold  | 99.12     | 80.00     | $F_{21} + F_{28} + (-F_8^2 + \cos(F_{29}))/((F_{22} + \tan(F_{24}))^2$                 |

$$TBF_{Musk1} = \frac{\exp(F_{128}^2)}{\cos(F_{117}) - \sin(F_{136})} - \frac{\sin(\exp(\sin(\tan(\sin(F_2) + \sin(F_{140}))))))}{\cos(\exp(\exp(\sin(F_{166}))) - \exp(F_{125})) * (F_3 + F_{36} + \sin(\cos(\cos(\sin(F_{19})))\sin(\sin(\exp(F_{44})))) - \cos(\exp(F_{45} - F_{166})) + \exp(F_{151}) + \tan(F_{147}))) * \frac{1}{\cos(F_{103} + F_{137}) + (F_{19} + F_{64} + F_{116})}} - \frac{\sin^2(F_{103})}{(F_{147} + F_{149})^2} - \frac{F_{57}^4 \exp(\sin(\exp(F_{142})))}{F_{77}^2 \tan^4(F_{36})} + \quad (13)$$

式(13)表示Musk1数据集在10-fold交叉验证方式下,10次独立重复实验中,最优分类准确率对应的数据特征及其对应类别之间的映射函数。

从表3可看出:FSLDGEP在92%的测试上 $TBCA$ 超过80%,在57%的测试上 $TBCA$ 超过90%。在71%的测试上 $TBDR$ 超过85%。在10-fold验证情况下,FSLDGEP针对Hepatitis、Ionosphere以及WDBC数据集,分别选择了3个、5个以及4个特征,构建出的分类器函数最优分类准确率达到100%,针对含有60个特征的Sonar数据集,FSLDGEP仅保留了3个特征,构建出了最佳分类准确度95.24%的分类函数;针对含有166个特征的Musk1数据集,FSLDGEP也仅仅是选择了使用22个特征去构造映射函数。对于大部分特征选择算法而言,它们仅仅反映了一种特征及其类别之间一种不明确的关系。但FSLDGEP可以从原始特征中选择少量的特征构造出与数据类别相关的有效映射函数关系式,其形式较为简单,仅通过几个初等函数组合得到,能更好地解释特征与类别之间的关系,虽然式(13)相对其他分类函数更为复杂,但由于Musk1数据拥有较多的特征,其类别由较多的特征决定,无

法通过少数特征与线性或非线性函数构造得到分类函数,本文算法在保证维度缩减率及分类准确率较好的条件下,从而得到了一个较为复杂的分类函数。尽管本文算法在其他数据集上有较好的表现,但对于Musk1数据集仍存在一点不足。通过最佳分类准确率及维度缩减率可以发现,最佳分类准确率与维度缩减率之间相差不大,大部分相差在10%左右,说明本文算法能够有效地平衡特征选择数量以及分类准确率之间的关系,在提高分类正确率的同时也能减少一定的特征选择的数量。

为验证本文算法构建的映射函数在数据分类上的有效性及优越性,将FSLDGEP和公开发表文献中的算法在相同的数据集和验证方式下的实验结果进行比较,每一次验证均重复10次,取其平均值。对比结果中:加粗数值表示在不同验证方式下的最优值;“—”表示该算法不存在该结果或无需采用该类方法策略。

各对比算法采用的分类器有J48、3近邻(3-Nearest Neighbor, 3NN)分类算法、1近邻(1-Nearest Neighbor, 1NN)分类算法、5近邻(5-Nearest Neighbor, 5NN)分类算法、基于径





向基核函数的支持向量机(SVM based on Radial basis kernel function, Rbf-SVM)、SVM、随机森林算法(Random Forests, RF)、决策树(Decision Tree, DT)等方法。

表 4 为 FSLDGEP 与森林优化特征选择算法(Feature Selection based on Forest Optimization Algorithm, FSFOA)<sup>[16]</sup>、基于邻域有效信息比的特征选择(Feature Selection based on the Neighborhood Effective Information Ratio, FS-NEIR)算法<sup>[25]</sup>、邻域软边界特征选择(feature evaluation and selection based on Neighborhood Soft Margin, NSM)<sup>[26]</sup>算法、基于离散互信息的分步优化特征选择算法(Step by step Optimization Feature Selection algorithm based on Discrete mutual information, SOFS-D)<sup>[5]</sup>、基于邻域互信息的分步优化特征选择算法(Step by step Optimization Feature Selection algorithm based on Neighborhood mutual information, SOFS-N)<sup>[5]</sup>、基于蚁群优化的无监督特征选择算法(Unsupervised Feature Selection algorithm based on Ant Colony Optimization, UFSACO)<sup>[19]</sup>等算法在 Hepatitis 数据集上的实验结果。

表 4 Hepatitis 数据集上的实验结果比较

Tab. 4 Comparison of experimental results on Hepatitis dataset

| 验证方式    | 算法      | 分类器 | CA/%         | DR/%         |
|---------|---------|-----|--------------|--------------|
| 10-fold | FSLDGEP | —   | <b>91.04</b> | <b>77.89</b> |
|         | FSFOA   | J48 | 86.45        | 55.00        |
|         | FS-NEIR | J48 | 81.11        | 68.42        |
|         | FSFOA   | 3NN | 87.09        | 42.10        |
|         | NSM     | 3NN | 90.00        | 15.78        |
|         | SOFS-D  | SVM | 84.07        | —            |
|         | SOFS-N  | SVM | 84.07        | —            |
| 70%-30% | FSLDGEP | —   | <b>86.30</b> | <b>82.11</b> |
|         | UFSACO  | J48 | 78.87        | 75.00        |
|         | FSFOA   | J48 | 84.40        | 45.00        |

表 5 为 FSLDGEP 与 FSFOA、新的森林优化算法的特征选择算法(New Feature Selection based on Forest Optimization Algorithm, NFSFOA)<sup>[2]</sup>、NSM、FS-NEIR、UFSACO、基于模糊互补准则的支持向量机特征选择方法(novel SVM-based feature selection method using a Fuzzy Complementary criterion, SVM-FuzCoc)<sup>[13]</sup>、基于离散域采样分类优化特征选择算法(Feature Selection algorithm using Sampling-And-Classification optimization algorithm, FSSAC)<sup>[17]</sup>、基于粒子群优化的特征选择方法分类的新初始化和更新机制(本文记为 PSO(4-2)), 4-2 表示原文作者采用自定义的第 4 种策略做种群初始化, 自定义的第 2 种策略更新种群<sup>[27]</sup>、基于互信息的包裹式特征选择混合遗传算法(Hybrid Genetic Algorithm for Feature Selection wrapper based on mutual information, HGAFS)<sup>[28]</sup>等算法在 Ionosphere 数据集上的实验结果。

表 6 为 FSLDGEP 与联合互信息特征选择算法(Joint Mutual Information feature selection algorithm, JMI)<sup>[29]</sup>、基于动态变化类的特征选择算法(Dynamic Change of Selected Feature algorithm with the class, DCSF)<sup>[29]</sup>、最小化冗余信息的动态特征选择方法(Dynamic Feature Selection method with Minimize Redundancy Information, MRIDFS)<sup>[29]</sup>、最小冗余与最大相关的特征选择算法(Minimal-Redundancy-Maximal-Relevance feature selection, MRMR)<sup>[30]</sup>、最大相关性和最大独立性特征选择算法(Max-Relevance and max-Independence

feature selection algorithm, MRI)<sup>[30]</sup>等在 Musk1 数据集上的实验结果。

表 5 Ionosphere 数据集上的实验结果比较

Tab. 5 Comparison of experimental results on Ionosphere dataset

| 验证方式    | 算法         | 分类器          | CA/%         | DR/%         |
|---------|------------|--------------|--------------|--------------|
| 10-fold | FSLDGEP    | —            | <b>94.87</b> | 84.12        |
|         | NFSFOA     | 3-NN         | 93.83        | 76.47        |
|         | NSM        | 3-NN         | 92.00        | <b>88.23</b> |
|         | NFSFOA     | 5-NN         | 93.23        | 79.41        |
|         | FSFOA      | J48          | 93.16        | 68.57        |
|         | FS-NEIR    | J48          | 92.59        | 82.35        |
| 70%-30% | FSLDGEP    | —            | 90.76        | 85.29        |
|         | SVM-FuzCoc | 1-NN         | 89.46        | 88.23        |
|         | NFSFOA     | 1-NN         | <b>95.16</b> | 61.76        |
|         | UFSACO     | J48          | 88.61        | 11.17        |
|         | FSFOA      | J48          | 95.12        | 47.05        |
|         | PSO(4-2)   | 5-NN         | 87.27        | <b>90.41</b> |
| 2-fold  | FSSAC      | Rbf-SVM/1-NN | 92.38        | 52.94        |
|         | FSLDGEP    | —            | 90.31        | <b>89.71</b> |
|         | NFSFOA     | 1-NN         | <b>95.16</b> | 58.82        |
|         | FSFOA      | Rbf-SVM      | 94.58        | 57.14        |
|         | HGAFS      | Rbf-SVM      | 92.76        | 82.35        |

表 6 Musk1 数据集上基于 10-fold 验证方式的实验结果比较

Tab. 6 Comparison of experimental results based on 10-fold verification on Musk1 dataset

| 算法      | 分类器 | CA/%         | DR/%         |
|---------|-----|--------------|--------------|
| FSLDGEP | —   | <b>75.20</b> | <b>91.99</b> |
| JMI     | SVM | 72.10        | —            |
| DCSF    | SVM | 72.89        | —            |
| MRIDFS  | SVM | 72.01        | —            |
| MRMR    | SVM | 70.75        | —            |
| MRI     | SVM | 69.71        | —            |

表 7 为 FSLDGEP 与基于特征子集区分度的顺序前向搜索特征选择算法(feature selection algorithm of Discernibility of Feature Subset based on Sequential Forward Search, DFS-SFS)<sup>[10]</sup>、基于特征相关性的顺序前向搜索特征选择算法(feature selection algorithm of Correlation of Feature Selector based on Sequential Forward Search, CFS-SFS)<sup>[10]</sup>、基于皮尔逊相关系数绝对值的顺序前向搜索相关特征选择(Correlation based Feature Selector based on the absolute of Pearson's correlation coefficient on Sequential Forward Search, CFSPabs-SFS)<sup>[10]</sup>、基于特征子集区分度的顺序后向搜索特征选择算法(feature selection algorithm of Discernibility of Feature Subset based on Sequential Backward Search, DFS-SBS)<sup>[10]</sup>、基于特征相关性的顺序后向搜索特征选择算法(feature selection algorithm of Correlation of Feature Selector based on Sequential Backward Search, CFS-SBS)<sup>[10]</sup>、基于皮尔逊相关系数绝对值的顺序后向搜索相关特征选择(Correlation based Feature Selector based on the absolute of Pearson's correlation coefficient on Sequential Backward Search, CFSPabs-SBS)<sup>[10]</sup>等算法在 WPBC 数据集上的实验结果。

表 8 为 FSLDGEP 与 FSFOA、FS-NEIR、基于高斯核粗糙集依赖性的特征选择算法(Feature Selection algorithm based on Dependency of Gaussian kernel rough set, FS-GD)<sup>[31]</sup>、基于邻域



粗糙集依赖性的特征选择算法 (Feature Selection algorithm based on Dependency of Neighborhood rough set, FS-ND)<sup>[32]</sup>、基于邻域识别率的特征选择算法 (Feature Selection algorithm based on Neighborhood Recognition Rate, FS-NRR)<sup>[33]</sup>、NFSFOA、粒子群优化 (Particle Swarm Optimization, PSO) 算法<sup>[27]</sup>、SVM-FuzCoc、改进的基于乌鸦搜索算法的特征选择算法 (Improved Feature Selection algorithm based on Crow Search Algorithm, IFSCrSA)<sup>[34]</sup>等在 Sonar 数据集上的实验结果。

表 7 WPBC 数据集上基于 5-fold 验证方式的实验结果比较

Tab. 7 Comparison of experimental results based on 5-fold verification on WPBC dataset

| 算法          | 分类器 | CA/%         | DR/%         |
|-------------|-----|--------------|--------------|
| FSLDGEP     | —   | <b>80.23</b> | <b>82.67</b> |
| DFS-SFS     | SVM | 76.29        | 2.00         |
| CFS-SFS     | SVM | 77.34        | 81.33        |
| CFSPabs-SFS | SVM | 75.78        | 79.33        |
| DFS-SBS     | SVM | 76.30        | 5.33         |
| CFS-SBS     | SVM | 76.30        | 72.67        |
| CFSPabs-SBS | SVM | 75.80        | 64.00        |

表 8 Sonar 数据集上的实验结果比较

Tab. 8 Comparison of experimental results on Sonar dataset

| 验证方式    | 算法         | 分类器     | CA/%         | DR/%         |
|---------|------------|---------|--------------|--------------|
| 10-fold | FSLDGEP    | —       | 82.17        | <b>91.83</b> |
|         | FSFOA      | 1-NN    | 74.60        | 56.67        |
|         | FS-NEIR    | DT      | 75.97        | 91.66        |
|         | FS-GD      | Rbf-SVM | 77.78        | 90.00        |
|         | FS-ND      | Rbf-SVM | 77.02        | 90.00        |
|         | FS-NRR     | Rbf-SVM | 79.58        | 91.67        |
|         | FSFOA      | J48     | 82.69        | 52.45        |
|         | NFSFOA     | J48     | <b>85.18</b> | 65.00        |
|         | FSLDGEP    | —       | 76.45        | <b>91.00</b> |
| 70%-30% | PSO        | —       | <b>78.64</b> | 22.04        |
|         | SVM-FuzCoc | 1-NN    | 73.17        | 68.33        |
|         | NFSFOA     | 1-NN    | 76.19        | 76.67        |
|         | NFSFOA     | 5-NN    | 74.60        | 68.33        |
| 2-fold  | FSLDGEP    | —       | 75.00        | <b>91.67</b> |
|         | NFSFOA     | SVM     | <b>75.94</b> | 78.33        |
|         | FSFOA      | SVM     | 72.11        | 63.33        |
|         | IFSCrSA    | Rbf-SVM | 69.23        | 91.67        |

如表 9 为 FSLDGEP 与 FSFOA、NSM、NFSFOA、FS-NEIR、UFSACO、IFSCrSA、融合 Shapley 值和粒子群优化算法的混合特征选择算法 (Shapley Value and Particle Swarm Optimization, SVPSO)<sup>[35]</sup>、新的 Shapely 值嵌入遗传算法 (novel Shapely Value Embedded Genetic Algorithm, SVEGA)<sup>[36]</sup>等在 Heart-Statlog 数据集上的实验结果。

如表 10 为 FSLDGEP 与 SOFS-D、SOFs-N、DFA-SFS、CFS-SFS、CFSPaBS-SFS、DFS-SBS、CFS-SBS、CFSPaBS-SBS、UFSACO、相关性冗余特征选择 (Relevance-Redundancy Feature Selection, RRFS) 算法<sup>[19]</sup>、融合蚁群算法和随机森林的特征选择方法 (feature selection method based on Ant Colony Optimization and Random Forest, ACORF)<sup>[18]</sup>等在 WDBC 数据集上的实验结果。

从表 4~10 中分类准确率的实验结果可以看出,关于 Hepatitis、Ionosphere、Musk1、WPBC、Heart-Statlog 和 WDBC 数

据集在 10-fold 条件验证下以及关于 WPBC 和 WDBC 数据集在 5-fold 条件验证下,本文算法均达到最好;但 FSLDGEP 在 Ionosphere、Sonar 和 Heart-Statlog 这三个数据集中,存在部分交叉验证的情况下,其分类准确率分别低于 NFSFOA、PSO 算法和 IFSCrSA 算法。虽然 FSLDGEP 不能保证在每一种验证条件下的分类准确率最高,但是 FSLDGEP 的分类准确率与数据集的当前最优分类准确率之间相差不大,并且在 10-fold 验证条件下,FSLDGEP 在大部分数据集上能够取得最好的分类准确率。

表 9 Heart-Statlog 数据集上的实验结果比较

Tab. 9 Comparison of experimental results on Heart-Statlog dataset

| 验证方式    | 算法      | 分类器     | CA/%         | DR/%         |
|---------|---------|---------|--------------|--------------|
| 10-fold | FSLDGEP | —       | <b>87.41</b> | 63.85        |
|         | FSFOA   | 3-NN    | 85.18        | 35.71        |
|         | NSM     | 3-NN    | 84.00        | 69.23        |
|         | NFSFOA  | 3-NN    | 83.33        | 53.85        |
|         | FSFOA   | J48     | 85.15        | 48.07        |
|         | FS-NEIR | J48     | 79.86        | 46.15        |
|         | UFSACO  | J48     | 81.48        | 16.15        |
|         | NFSFOA  | J48     | 84.07        | 61.54        |
|         | UFSACO  | Rbf-SVM | 72.22        | 53.85        |
|         | IFSCrSA | Rbf-SVM | 85.56        | 50.00        |
|         | NSM     | SVM     | 86.00        | <b>76.92</b> |
|         | SVEGA   | SVM     | 84.88        | 42.86        |
|         | SVPSO   | SVM     | 85.71        | 57.14        |
|         | FSLDGEP | —       | 80.74        | 65.38        |
| 2-fold  | FSFOA   | SVM     | 84.07        | 50.00        |
|         | NFSFOA  | SVM     | 84.81        | <b>76.92</b> |
|         | IFSCrSA | 3-NN    | <b>85.89</b> | 72.58        |

表 10 WDBC 数据集上的实验结果比较

Tab. 10 Comparison of experimental results on WDBC dataset

| 验证方式    | 算法          | 分类器 | CA/%         | DR/%         |
|---------|-------------|-----|--------------|--------------|
| 10-fold | FSLDGEP     | —   | <b>97.19</b> | <b>83.00</b> |
|         | SOFS-D      | SVM | 95.78        | —            |
|         | SOFS-N      | SVM | 94.61        | —            |
|         | SOFS-D      | J48 | 94.38        | —            |
|         | SOFS-N      | J48 | 94.55        | —            |
| 5-fold  | FSLDGEP     | —   | <b>97.72</b> | <b>84.67</b> |
|         | DFS-SFS     | SVM | 64.09        | 6.67         |
|         | CFS-SFS     | SVM | 63.09        | 53.33        |
|         | CFSPaBS-SFS | SVM | 63.09        | 50.00        |
|         | DFS-SBS     | SVM | 64.68        | 10.00        |
|         | CFS-SBS     | SVM | 63.61        | 62.00        |
|         | CFSPaBS-SBS | SVM | 64.69        | 18.67        |
| —       | UFSACO      | SVM | 90.72        | 83.33        |
|         | RRFS        | SVM | 90.36        | 83.33        |
|         | ACORF       | RF  | 96.00        | 83.33        |

从表 4~10 中维度缩减率的实验结果可以看出,关于 Hepatitis、WPBC、Sonar 和 WDBC 数据集在几种不同验证条件下,本文算法均达到最优。但在 Ionosphere 数据集上基于 10-fold 与 70%-30% 验证条件下分别低于 NSM 算法与 PSO (4-2) 算法;在 Heart-Statlog 数据集上基于 10-fold 与 2-fold 验证条件下分别低于 NSM 算法与 NFSFOA 算法。尽管维度缩减率相较于上述算法略有不足,但 FSLDGEP 在大部分数据集上的不同验证方法下的维度缩减率超过了 80%,只有





Hepatitis 和 Heart-Statlog 数据集基于 10-fold 验证条件下的维度缩减率分别为 77.89% 和 65.38%。出现这样的情况并不能说明本文算法在 10-fold 验证条件下不适用于 Hepatitis 和 Heart-Statlog 数据集;相反地,利用本文算法对这两个数据集进行特征选择分类后,均能达到该数据集在 10-fold 验证条件下的最高分类准确率。

综合表 4~10 中的 CA 与 DR 的实验结果可以发现, FSLDGEP 在 Musk1、WPBC、WDBC 这 3 个数据集上均取得了最好结果。尽管在有些数据集的不同验证条件下, FSLDGEP 不能使得 CA 与 DR 同时取得最优值,但是, FSLDGEP 在大部分情况下能够使得 CA 与 DR 形成一种互补的情况。在 Ionosphere 数据集上,基于 10-fold 验证条件下, FSLDGEP 与 NSM 算法相比,在 DR 上低 4.11 个百分点,但 CA 上高 2.87 个百分点;基于 70%-30% 与 2-fold 验证条件下, FSLDGEP 与 NFSFOA 算法相比,在 CA 上分别低 4.4 与 4.85 个百分点,但在 DR 上高 23.53 与 30.89 个百分点。在 Sonar 数据集上,基于 10-fold 与 2-fold 验证条件下, FSLDGEP 与 NFSFOA 相比,在 CA 上低 3.01 与 0.94 个百分点,但在 DR 上高 26.83 与 13.34 个百分点;基于 70%-30% 验证条件下, FSLDGEP 与 PSO 算法相比,在 CA 上低 2.19 个百分点,但 DR 上高 68.96 个百分点;从 CA 与 DR 的结果也可以反映出, FSLDGEP 相较于其他算法来说占有一定的优势。

综合表 3~10 的结果可以看出,无论是低维度特征数据 Heart-Statlog,还是高纬度特征数据 Musk1, FSLDGEP 都能够很好地发现数据特征及其所属类别之间的一个函数映射关系,且发现的函数能很好地平衡 CA 与 DR。说明了本文算法生成的映射函数在数据特征提取分类上的可行性与有效性,同时说明了本文算法在数据特征选择分类上的优越性。

#### 4 结语

本文针对特征选择算法无法揭示特征与其类别之间存在的具体的关系的问题,提出了一种改进的层次距离基因表达式编程特征选择分类算法。该方法首先依概率对种群个体进行初始化,使个体头部侧重于函数符号的选择,减少无效基因的表达,提高了种群进化的速度;其次,通过层次距离确定遗传位点,并对其进行遗传操作,利用定义的层次邻域,使得种群个体的层次变异具有导向性,提升了算法寻优的精度;最后,引入新的适应度评价指标,综合考量了分类准确率和维度缩减率,改变了种群更新机制的局限性,使得算法能够更有效地建立特征与类别之间的函数关系,为特征的选择提供了一种新的思路。在 7 个数据集上进行仿真实验,与 SVM、NN 等分类器相比, FSLDGEP 在能够有效构建特征及其类别之间函数关系式,减少特征选取的数量,提升分类效果。针对高维特征数据,如何保证高分类率和低特征选择率的同时,仍能得到一个较为简单的函数表达式,将是进一步的研究工作。

#### 参考文献 (References)

- [1] 张翠军,陈贝贝,周冲,等. 基于多目标骨架粒子群优化的特征选择算法[J]. 计算机应用, 2018, 38(11): 3156-3160, 3166. (ZHANG C J, CHEN B B, ZHOU C, et al. Feature selection algorithm based on multi-objective bare-bones particle swarm optimization [J]. Journal of Computer Applications, 2018, 38(11): 3156-3160, 3166.)
- [2] 谢琪,徐旭,程耕国,等. 基于新的森林优化算法的特征选择算法[J]. 计算机应用, 2020, 40(5): 1266-1271. (XIE Q, XU X, CHENG G G, et al. Feature selection algorithm based on new forest optimization algorithm [J]. Journal of Computer Applications, 2020, 40(5): 1266-1271.)
- [3] 谢娟英,王明钊,周颖,等. 非平衡基因数据的差异表达基因选择算法研究[J]. 计算机学报, 2019, 42(6): 1232-1251. (XIE J Y, WANG M Z, ZHOU Y, et al. Differential expression gene selection algorithms for unbalanced gene datasets [J]. Chinese Journal of Computers, 2019, 42(6): 1232-1251.)
- [4] 王翔,胡学钢. 高维小样本分类问题中特征选择研究综述[J]. 计算机应用, 2017, 37(9): 2433-2438, 2448. (WANG X, HU X G. Overview on feature selection in high-dimensional and small-sample-size classification [J]. Journal of Computer Applications, 2017, 37(9): 2433-2438, 2448.)
- [5] 樊鑫,陈红梅. 基于差别矩阵和 mRMR 的分步优化特征选择算法[J]. 计算机科学, 2020, 47(1): 87-95. (FAN X, CHEN H M. Stepwise optimized feature selection algorithm based on discernibility matrix and mRMR [J]. Computer Science, 2020, 47(1): 87-95.)
- [6] 姚登举. 面向医学数据的随机森林特征选择及分类方法研究[D]. 哈尔滨: 哈尔滨工程大学, 2016: 1-2. (YAO D J. Research on feature selection and classification method based on random forest for medical datasets [D]. Harbin: Harbin Engineering University, 2016: 1-2.)
- [7] 李占山,刘兆康. 基于 XGBoost 的特征选择算法[J]. 通信学报, 2019, 40(10): 101-108. (LI Z S, LIU Z G. Feature selection algorithm based on XGBoost [J]. Journal on Communications, 2019, 40(10): 101-108.)
- [8] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods[J]. Computers and Electrical Engineering, 2014, 40(1): 16-28.
- [9] SHARMIN S, SHOYAIB M, ALI A A, et al. Simultaneous feature selection and discretization based on mutual information [J]. Pattern Recognition, 2019, 91: 162-174.
- [10] 谢娟英,谢维信. 基于特征子集区分度与支持向量机的特征选择算法[J]. 计算机学报, 2014, 37(8): 1704-1718. (XIE J Y, XIE W X. Several feature selection algorithms based on the discernibility of a feature subset and support vector machines [J]. Chinese Journal of Computers, 2014, 37(8): 1704-1718.)
- [11] 张鑫. 基于自然进化策略的特征选择算法研究[D]. 长春: 吉林大学, 2020: 1-2. (ZHANG X. Research of feature selection algorithm based on natural evolution strategy [D]. Changchun: Jilin University, 2020: 1-2.)
- [12] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 72-75. (LI H. Statistical Learning Methods [M]. Beijing: Tsinghua University Press, 2012: 72-75.)
- [13] MOUSTAKIDIS S P, THEOCHARIS J B. SVM-FuzCoC: a novel SVM-based feature selection method using a fuzzy complementary criterion[J]. Pattern Recognition, 2010, 43(11): 3712-3729.
- [14] SUN Z Q, ZHANG J, DAI L, et al. Mutual information based multi-label feature selection via constrained convex optimization [J]. Neurocomputing, 2019, 329: 447-456.
- [15] SHEIKHI G, ALTINÇAY H. maximum-relevance and maximum-diversity of positive ranks: a novel feature selection method [J]. Expert Systems with Applications, 2020, 158: No. 113499.
- [16] GHAEMI M, FEIZI-DERAKHSHI M R. Feature selection using forest optimization algorithm [J]. Pattern Recognition, 2016, 60:



- 121-129.
- [17] 张梦林,李占山. 基于SAC的特征选择算法[J]. 计算机科学, 2018, 45(2): 63-68. (ZHANG M L, LI Z S. Feature selection algorithm using SAC algorithm[J]. Computer Science, 2018, 45(2): 63-68.)
- [18] 李光华,李俊清,张亮,等. 一种融合蚁群算法和随机森林的特征选择方法[J]. 计算机科学, 2019, 46(11A): 212-215. (LI G H, LI J Q, ZHANG L, et al. Feature selection method based on ant colony optimization and random forest[J]. Computer Science, 2019, 46(11A): 212-215.)
- [19] TABAKHI S, MORADI P, AKHLAGHIAN F. An unsupervised feature selection algorithm based on ant colony optimization[J]. Engineering Applications of Artificial Intelligence, 2014, 32: 112-123.
- [20] MA J B, GAO X Y. Designing genetic programming classifiers with feature selection and feature construction[J]. Applied Soft Computing, 2020, 97(Pt B): No. 106826.
- [21] MA J B, GAO X Y. A filter-based feature construction and feature selection approach for classification using Genetic Programming[J]. Knowledge-Based Systems, 2020, 196: No. 105806.
- [22] FERREIRA C. Gene expression programming: a new adaptive algorithm for solving problems[J]. Complex Systems, 2001, 13(2): 87-129.
- [23] 崔未. 基于层次距离的GEP算法及其应用[D]. 武汉:武汉理工大学, 2019: 17-28. (CUI W. Application of gene expression programming based on layer distance [D]. Wuhan: Wuhan University of Technology, 2019: 17-28.)
- [24] DUA D, GRAFF C. UCI machine learning repository [DS/OL]. [2020-11-11]. <http://archive.ics.uci.edu/ml>.
- [25] ZHU W Z, SI G Q, ZHANG Y B, et al. Neighborhood effective information ratio for hybrid feature subset evaluation and selection[J]. Neurocomputing, 2013, 99: 25-37.
- [26] HU Q H, CHE X J, ZHANG L, et al. Feature evaluation and selection based on neighborhood soft margin[J]. Neurocomputing, 2010, 73(10/11/12): 2114-2124.
- [27] XUE B, ZHANG M J, BROWNE W N. Particle swarm optimisation for feature selection in classification: novel initialisation and updating mechanisms [J]. Applied Soft Computing, 2014, 18: 261-276.
- [28] HUANG J J, CAI Y Z, XU X M. A hybrid genetic algorithm for feature selection wrapper based on mutual information[J]. Pattern Recognition Letters, 2007, 28(13): 1825-1844.
- [29] ZHOU H F, WEN J. Dynamic feature selection method with minimum redundancy information for linear data [J]. Applied Intelligence, 2020, 50(11): 3660-3677.
- [30] GAO W F, HU L, ZHANG P. Class-specific mutual information variation for feature selection[J]. Pattern Recognition, 2018, 79: 328-339.
- [31] HU Q H, ZHANG L, CHEN D G, et al. Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications [J]. International Journal of Approximate Reasoning, 2010, 51(4): 453-471.
- [32] HU Q H, YU D R, LIU J F, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences, 2008, 178(18): 3577-3594.
- [33] HU Q H, PEDRYCZ W, YU D R, et al. Selecting discrete and continuous features based on neighborhood decision error minimization [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2010, 40(1): 137-150.
- [34] 王颖,曹捷,邱志洋. 基于乌鸦搜索算法的新型特征选择算法[J]. 吉林大学学报(理学版), 2019, 57(4): 869-874. (WANG Y, CAO J, QIU Z Y. A novel feature selection algorithm based on crow search algorithm [J]. Journal of Jilin University (Science Edition), 2019, 57(4): 869-874.)
- [35] 邓秀勤,李文洲,武继刚,等. 融合Shapley值和粒子群优化算法的混合特征选择算法[J]. 计算机应用, 2018, 38(5): 1245-1249. (DENG X Q, LI W Z, WU J G, et al. Hybrid feature selection algorithm fused Shapley value and particle swarm optimization [J]. Journal of Computer Applications, 2018, 38(5): 1245-1249.)
- [36] SASIKALA S, APPAVU ALIAS BALAMURUGAN S, GEETHA S. A novel adaptive feature selector for supervised classification [J]. Information Processing Letters, 2016, 117: 25-34.

This work is partially supported by the Surface Program of National Natural Science Foundation of China (61672391).

**ZHAN Hang**, born in 1996, M. S. candidate. His research interests include intelligent computing.

**HE Lang**, born in 1974, Ph. D., professor. His research interests include evolutionary computation, intelligent computing.

**HUANG Zhangcan**, born in 1960, Ph. D., professor. His research interests include intelligent computing, image processing.

**LI Huafeng**, born in 1995, M. S. candidate. His research interests include intelligent computing.

**ZHANG Qiang**, born in 1996, M. S. candidate. Her research interests include intelligent computing.

**TAN Qing**, born in 1991, Ph. D. candidate. His research interests include intelligent computing, image processing.