CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 46 卷第 2 期 2024 年 2 月
Vol. 46, No. 2, Feb. 2024

文章编号:1007-130X(2024)02-0282-10

融合特征权重与改进粒子群优化的特征选择算法*

刘振超[1],苑迎春[1,2],王克俭[1,2],何　晨[1]

(1.河北农业大学信息科学与技术学院,河北 保定 071000;
2.河北农业大学河北省农业大数据重点实验室,河北 保定 071000)

摘　要:随着教育信息化的发展,教育数据呈现特征数量高、冗余度高等特点,这使目前的分类算法在教育数据上分类准确率不理想。提出一种将特征权重算法与改进粒子群优化算法融合的混合式特征选择算法(RF-ATPSO)。该算法首先使用 RELIEF-F 算法计算各个特征的权重,筛除冗余特征,然后在筛选后的特征集合中利用改进粒子群算法搜索最优特征子集。实验结果表明,在 6 个 UCI 公共数据集上,经 RF-ATPSO 算法进行特征选择后,平均准确率提升了 10.04%,且平均特征子集规模最小、收敛速度最快;在学生学业成绩画像特征数据集上,该算法以较小的特征子集规模达到较高的分类准确率,平均准确率为 94.77%,明显优于其它特征选择算法,实验充分证明了该算法具有实际应用意义。

关键词:特征选择;特征权重;改进粒子群优化;T-分布

中图分类号:TP391　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2024.02.011

Feature selection algorithm based on feature weights and improved particle swarm optimization

LIU Zhen-chao[1], YUAN Ying-chun[1,2], WANG Ke-jian[1,2], HE Chen[1]

(1. College of Information Science and Technology, Hebei Agricultural University, Baoding 071000;
2. Hebei Key Laboratory of Agricultural Big Data, Hebei Agricultural University, Baoding 071000, China)

Abstract: With the development of educational informatization, educational data presents characteristics such as high feature counts and high redundancy, resulting in the classification accuracy of current classification algorithms not being ideal on educational data. Therefore, this paper proposes a hybrid feature selection algorithm (RF-ATPSO) that integrates feature weighting algorithm with improved particle swarm optimization algorithm. The algorithm first uses the RELIEF-F algorithm to calculate the weights of each feature, removes redundant features, and then uses the improved particle swarm optimization algorithm to search for the optimal feature subset in the filtered feature set. Experimental results show that on 6 UCI public datasets, after feature selection using the RF-ATPSO algorithm, the average accuracy is improved by 10.04%, and the average feature subset size is the smallest and the convergence speed is the fastest. In the student academic performance portrait feature dataset, the algorithm achieves high classification accuracy with a smaller feature subset size, with an average accuracy of 94.77%, which is significantly better than other feature selection algorithms. The experiment fully demonstrates the practical application significance of this algorithm.

Key words: feature selection; feature weight; improved PSO; T-distribution

* **收稿日期**:2022-08-23;**修回日期**:2022-10-24
基金项目:河北省高等教育教学改革研究与实践项目(2020GJJG076)
通信作者:苑迎春(nd_hd_yyc@163.com)
通信地址:071000 河北省保定市河北农业大学信息科学与技术学院
Address: College of Information Science and Technology, Hebei Agricultural University, Baoding 071000, Hebei, P. R. China

1 引言

大数据时代,高等教育信息建设迅速发展,使得高校的教育教学数据逐年剧增,有效挖掘并合理利用高校教学数据对学校管理、教师教学及学生自我认知的提升都具有重要价值[1]。决策树因其分类准确率高、运算效率高,被广泛应用于教育教学数据挖掘中。但是,由于教育教学等数据具有维度高、冗余多等特点,若将高维原始数据直接应用于决策树分类,决策树分类的准确率并不理想。特征选择[2]是数据预处理的关键步骤,是从原始特征集合中筛选出对分类模型性能贡献度最高的特征子集。特征选择不但能有效降低数据集特征维度,提升分类模型的学习效率,还可以从原始数据集中选择对分类器分类性能贡献最高的特征子集,从而提高分类器的分类准确率[3]。常见的特征选择算法根据其是否包含相关学习算法可以分为过滤式(Filter)和封装式(Wrapper)2 种。

Filter 特征选择算法[4]通过数据非标签特征与标签特征之间的潜在规律以及数据本身内在性质判断数据特征的优劣,进而筛选特征子集。常用方法有互信息法[5]、信息增益法[6]和特征权重法[7]等。该类算法具有简单易行、效率较高和评价标准独立于分类算法等特点。

Wrapper 特征选择算法由分类器和搜索算法组成,以分类器的分类准确率作为性能评估标准,通过对原始数据特征集进行搜索得到特征子集。该类算法能有效筛除冗余特征,提高分类准确率。已有研究人员利用粒子群优化 PSO(Particle Swarm Optimization)算法[8]、灰狼算法 GWO(Grey Wolf Optimizer)[9]等元启发式算法作为搜索策略,有效提高了所选特征质量和数据分类准确率。例如,吴晓燕等[10]利用樽海鞘群算法和粒子群优化 PSO 算法进行特征选择,在不同 UCI(University of California, Irvine)数据集上均可选出最佳特征子集,并在多项评估指标上获得了较好效果;Zhang 等[11]提出利用粒子群搜索特征子集的封装式算法,使用 C4.5 算法作为评估算法,实验结果表明该算法提取的特征子集有较高的辨识度。Wrapper 算法尽管提升了分类准确率,但当数据维度较高时,仍存在计算代价高、效率低等问题。

Filter 和 Wrapper 特征选择算法在特征选择方面各有优势和不足,因此有研究人员[12]提出了将 2 类算法融合使用的特征选择策略。该策略的一般流程为:首先,使用 Filter 算法剔除部分冗余特征,以减小启发式算法的特征搜索规模;然后,将 Filter 算法筛选出的特征子集传递给 Wrapper 算法,再进一步搜索最优特征子集。王金杰等[13]将粒子群优化算法和互信息融合成混合式多目标特征选择方法,在 15 个 UCI 数据集上的实验结果表明,该算法能够有效减少特征个数,降低分类错误率。肖艳等[14]针对面向对象土地分类中数据特征维数过高的问题,提出了将 RELIEF-F 和粒子群优化算法混合的特征选择算法,有效降低了土地数据维度,提高了面向对象土地分类的效率。虽然上述文献均使用了包含粒子群的融合式特征选择算法,在特征选择方面进行了有效改进,但相关研究表明[15,16]:粒子群优化算法可能因迭代初期种群个体多样性的快速降低使得算法收敛过早,出现“早熟”现象,进而影响特征选择算法的性能。

综上分析,本文提出一种融合特征权重与改进粒子群优化算法的混合式特征选择算法 RF-ATPSO(RELIEF-F Adaptive T-distribution Particle Swarm Optimization)。该算法利用特征权重过滤法剔除部分冗余特征,有效降低后续改进粒子群优化算法的搜索规模;通过自适应权重和 T-分布扰动 2 种改进策略,平衡粒子群优化算法的全局探索和局部开发能力,提高粒子群的多样性,进而保证在 Wrapper 算法特征选择时不易陷入局部最优,从而提高算法的特征选择性能。

2 改进粒子群优化算法

搜索算法是 Wrapper 特征选择算法中的关键组成部分;而粒子群优化 PSO 算法因其优越的全局搜索和寻优能力在各个领域被广泛应用。因此,本文利用粒子群优化算法在 Wrapper 算法中搜索最优特征子集。但其迭代初期种群个体多样性的快速降低会使得算法收敛过早,容易陷入局部最优,进而影响选出的特征子集的质量。因此,本节通过自适应惯性权重和 T-分布扰动 2 种策略改进粒子群优化算法,提高其寻优能力,从而提高 Wrapper 特征选择算法的性能。

2.1 粒子群优化算法基本原理

粒子群优化算法是 Kennedy 等[17]根据鸟群捕食行为中寻找最佳觅食区域的过程所提出的一种智能算法,具有原理简单、参数少等优点。在粒子群优化算法中,鸟群中的每个个体都是一个粒子,

每个粒子均记录自己所找到的最佳觅食位置(局部最优解),粒子群中所有粒子的最佳觅食位置可以看作全局最优解,每个粒子的觅食位置拥有食物的可能性通过适应度刻画。

假设个体数为 N 的粒子群在 D 维空间中寻找最优解,其中第 i 个粒子在 N 维空间中可用位置 $\boldsymbol{X}_i=(x_{i,1},x_{i,2},\cdots,x_{i,D})$ 表示,第 i 个粒子的飞行速度设为 $\boldsymbol{V}_i=(v_{i,1},v_{i,2},\cdots,v_{i,D})$,第 i 个粒子的历史最优位置称为个体最优值 $\boldsymbol{P}_i=(p_{i,1},p_{i,2},\cdots,p_{i,D})$,整个粒子群的最优位置称为全局最优值 $\boldsymbol{G}_{\text{best}}=(g_{\text{best},1},g_{\text{best},2},\cdots,g_{\text{best},D})$。根据第 i 个粒子、第 i 个粒子最优值 $\boldsymbol{P}_i$ 和全局最优值 $\boldsymbol{G}_{\text{best}}$ 对粒子的速度和位置进行更新,更新公式如式(1)和式(2)所示:

$$v'_{i,d}=\omega\times v_{i,d}+c_1r_1(p_{i,d}-x_{i,d})+c_2r_2(g_{\text{best},d}-x_{i,d}) \tag{1}$$

$$x'_{i,d}=x_{i,d}+v_{i,d} \tag{2}$$

其中,ω 表示粒子的惯性权重,该值将会影响算法的收敛性;c_1 和 c_2 表示学习因子,即加速常数;r_1、r_2 表示 0～1 之间的随机数;$1\leqslant d\leqslant D$。

根据上述公式可以看出,粒子群优化算法寻优基于本身(局部最优)及周围个体的经验(全局最优)进行决策。在迭代初期,粒子群的个体多样性迅速降低,导致算法提前收敛,从而丢失一些重要的位置信息。针对以上不足,本文从 2 个方面对粒子群优化算法进行改进,平衡算法的全局探索和局部开发能力,提升粒子群优化算法的搜索精度。

2.2 自适应惯性权重策略

ω 为粒子的惯性权重,其取值将影响算法收敛性。在粒子群迭代过程中,算法迭代前期需要增加粒子变化步长,从而较早定位全局最优解所在的区域;算法迭代后期则需要减小粒子变化步长,使粒子在该区域内精细化搜索,以找到全局最优解。基于上述思想,本文提出自适应惯性权重策略来平衡算法的全局探索和局部开发能力。ω 的计算可用式(3)表示:

$$\omega=0.8\times e^{-3(t/t_{\max})^2} \tag{3}$$

其中,t 表示迭代次数,$t_{\max}$ 表示最大迭代次数。ω 在迭代初期尽可能取最大值,使算法步长迅速变化,方便进行全局搜索;随着迭代的进行,权重不断减小,侧重进行局部搜索。该策略有效平衡了算法的全局探索和局部开发能力。

2.3 T-分布扰动策略

在迭代初期粒子种群个体多样性迅速下降,导致迭代后期种群多样性较低。粒子群的群体最优值远离全局最优值时,粒子易向错误方向进化和学习,此情况下极易陷入局部最优。本文提出了一种基于 T-分布的扰动策略,以实现在算法迭代过程中增加粒子种群的多样性并及时跳出局部最优。即如果经过连续几次迭代,当前粒子的最优适应度值基本没有或不再发生变化,则认为算法陷入局部最优,在这时加入扰动让粒子震荡,使其跳出局部最优,这样也增加了种群多样性。该策略如式(4)所示:

$$\boldsymbol{X}_i^t=\boldsymbol{X}_i+\boldsymbol{X}_i\times h(t) \tag{4}$$

其中,$\boldsymbol{X}_i^t$ 表示扰动后的粒子位置;$h(t)$ 表示以当前迭代次数 t 为自由度的 T-分布的值。

3 融合特征权重和 ATPSO 的特征选择(RF-ATPSO)算法

在过滤式算法中,特征权重算法 RELIEF-F 具有运行效率高、特征选择结果辨识度好的优势。本文提出双策略改进粒子群优化算法平衡了全局探索和局部开发能力,增加了粒子的多样性,提高了粒子群优化算法的搜索能力。基于此,提出一种将特征权重算法 RELIEF-F 与改进粒子群优化算法融合的混合特征选择算法。该算法主要包括 2 部分:首先使用特征权重算法对原始特征集合进行初步特征筛选;然后从筛选后的特征集合中利用改进粒子群优化算法搜索最优特征子集,提高所选特征子集的精度及后来的分类准确率。其中又包括 2 个关键步骤,分别是粒子群二进制转化和适应度函数设计。

3.1 特征权重算法 RELIEF-F

RELIEF-F 算法是 Kononenko 等[18]在 1994 年基于 RELIEF 算法改进的一种适用于多分类的特征选择方法。

特征权重计算流程如下:

重复执行步骤(1)～步骤(3)共 m 次:

(1)从数据集中随机抽取样本 $\boldsymbol{R}$,选择 $\boldsymbol{R}$ 的猜中近邻和猜错近邻各 k 个,分别记作集合 $H=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_k\}$,$M=\{\boldsymbol{m}_1,\boldsymbol{m}_2,\cdots,\boldsymbol{m}_k\}$。

(2)根据以下规则进行特征权重更新:若 $\boldsymbol{R}$ 和 H 中所有样本在某个特征上的距离小于 $\boldsymbol{R}$ 和 M 中所有样本的距离,说明该特征对区分同类和异类样本最近邻有益,则增加该特征权重,反之降低该特征权重。

(3)根据式(5)和式(6)更新特征 A 的特征权重,直到最大迭代次数结束。

$$diff(A,\boldsymbol{R},\boldsymbol{h}_j)=\frac{|\boldsymbol{R}[A]-\boldsymbol{h}_j[A]|}{\max(A)-\min(A)} \tag{5}$$

$$W'(A)=W(A)-\sum_{j=1}^{k}diff(A,\boldsymbol{R},\boldsymbol{h}_j)/(m\times k)+\sum_{C\neq class(\boldsymbol{R})}\left[\frac{P(C)}{1-P(class(\boldsymbol{R}))}\sum_{j=1}^{k}diff(A,\boldsymbol{R},\boldsymbol{m}_j)\right]/(m\times k) \tag{6}$$

其中,A 表示样本的一种特征,$\max(A)$和 $\min(A)$分别表示特征 A 上的最大取值和最小取值,$\boldsymbol{R}[A]$表示样本 $\boldsymbol{R}$ 的特征 A 上的值,$\boldsymbol{h}_j[A]$表示猜中近邻中第 j 个样本 $\boldsymbol{h}_j$ 在特征 A 上的值;$diff(A,\boldsymbol{R},\boldsymbol{h}_j)$表示样本 $\boldsymbol{R}$ 与样本 $\boldsymbol{h}_j$ 在特征 A 上的差;$P(C)$表示 C 类的比例;$P(class(\boldsymbol{R}))$表示随机抽取样本 $\boldsymbol{R}$ 所属类别的比例;$\boldsymbol{m}_j$ 表示 C 类样本中的第 j 个最近邻样本。

3.2 粒子群二进制转化方式

基于 ATPSO(Adaptive T-distribution Particle Swarm Optimization)法对数据集进行特征选择,可以看作将解空间限定在{0,1}范围内的二进制优化问题。需要注意的一点是,进行特征选择时,需要将连续型优化问题转换为离散型优化问题。

首先要对粒子群中的粒子进行编码。一个完整的特征选择解对应改进粒子群优化算法中的一个粒子,粒子的维度与原始数据集中样本的特征属性数量相同,且粒子群个体的某个维度值 $x_{i,j}\in\{0,1\}$。若要将离散粒子群与特征选择问题正确对应,需定义粒子群编码规则。编码规则为:若 $x_{i,j}=1$,表明第 i 个粒子的第 j 个特征被选择,若 $x_{i,j}=0$,则表明第 i 个粒子的第 j 个特征未被选择。

除粒子编码问题外,连续型优化问题如何转换为离散型优化问题也同样重要。本文利用 Sigmoid 函数将连续型变量转换为二进制形式。Sigmoid 函数如式(7)所示:

$$f(x)=\frac{1}{1+e^{-x}} \tag{7}$$

具体到特征选择上,需要将连续型粒子的各个维度映射到{0,1},需将 $x_{i,j}$ 带入 Sigmoid 函数,结果如式(8)所示:

$$T(x_{i,j})=\frac{1}{1+e^{-x_{i,j}}} \tag{8}$$

其中,映射函数 $T(\cdot)$ 表示粒子中的元素 $x_{i,j}$ 取值为 1 的概率。综上所述,粒子群的位置更新策略可以用式(9)进行描述:

$$x_{i,j}=\begin{cases}1, rand\geqslant T(x_{i,j})\\0, rand\leqslant T(x_{i,j})\end{cases} \tag{9}$$

其中,$rand$ 为[0,1]的随机数。若随机数大于或等于元素 $x_{i,j}$ 取值为 1 的概率,则 $rand$ 取值为 1,否则取值为 0。

以粒子群的某一种特征选择解为例。假设原始数据集拥有 7 个特征,在 ATPSO 算法迭代中某个粒子位置的结果如图 1 所示。由图 1 可知,$x_{i,2}=x_{i,3}=x_{i,5}=x_{i,6}=1$,$x_{i,1}=x_{i,4}=0$,表明第 i 个粒子将原始数据特征 2,3,5 和 6 选中作为特征选择的最优特征子集,将原始数据特征 1 和 4 筛除。最终利用分类器可以基于选出来的最优特征子集进行模型训练与数据分类。

Figure 1 Feature selection solution

图 1 特征选择解

3.3 适应度函数设计

数据集的特征选择可以转化成多目标优化问题。优化目标为:在满足特征选择数量最小化的同时,也最大化分类器的分类准确率。基于上述 2 个优化目标,本文将适应度函数定义为式(10):

$$fitness=\alpha\times error\_rate+\beta\times\frac{RF}{D} \tag{10}$$

其中,$error\_rate$ 表示指定分类算法(本文采用决策树算法)的误分率,D 表示数据集中样本的特征总数量,RF 表示特征选择算法最终所选择的特征子集大小,α、β 分别对应分类算法误分率和特征子集大小在适应度中的重要性。α、$\beta\in[0,1]$,且 $\alpha+\beta=1$。

3.4 RF-ATPSO 算法流程

RF-ATPSO 特征选择算法首先使用特征权重算法对原始特征集合进行初步特征筛选,然后从筛选后的特征集合中利用改进粒子群优化算法搜索最优特征子集,最终得到最优特征子集。算法详细步骤如下所示:

算法 1 RF-ATPSO 特征选择算法

输入:基准数据集。

输出:最优特征子集以及分类算法的准确率。

步骤 1 输入基准数据集,将其按照 7∶3 的比例划分为训练集和测试集,设置 C4.5 为评估算法。

步骤 2 使用 RELIEF-F 算法计算各个特征权重并按照权重对特征排序。

步骤 3 根据设定阈值对有序的特征集进行筛选。

步骤 4 初始化粒子群优化算法参数,初始化粒子初始位置并利用式(9)实现粒子位置和特征集的映射。

步骤 5 利用式(10)计算粒子适应度值。

步骤 6 比较每个粒子的适应度值,更新全局和局部最优解。

步骤 7 利用自适应惯性权重策略(式(1)和式(2)所示),更新粒子位置。

步骤 8 执行 T-分布策略。

步骤 9 若未达到最大迭代次数则跳转至步骤 5。

步骤 10 输出最优特征子集和分类准确率。

4 实验与结果分析

4.1 数据集介绍与实验设置

4.1.1 数据集介绍

为充分验证本文提出的 RF-ATPSO 算法的有效性,本文基于加州大学 UCI 机器学习库中的 6 个标准数据集进行实验。这些数据集分别来自不同领域,如 Spambase 主要用于冗余邮件的识别分类,Arrhythmia 心率失常数据集和 Cancer 癌症数据集为医学数据集。

表 1 简要介绍了上述 6 个 UCI 数据集和学生画像指标数据集的样本数量、特征数量和类别数量。

Table 1 Datasets introduction

表 1 数据集介绍

数据集名称	样本数量	特征数量	类别数量
Meu	2 856	71	56
Scadi	70	205	7
Spambase	4 601	57	2
Cancer	32	56	3
Arrhythmia	452	279	2
HillValley	606	100	2
Stu_portrait	142	281	3

为进一步验证本文算法的鲁棒性,实验选用本研究团队构建的某高校学生学业成绩画像指标数据集。该数据集从学分体系模块、成绩体系模块和课程指标体系模块 3 个方面构建学业指标体系,全方位刻画学期、学年和课程类别等方面的学生学业成绩情况。

构建的学生学业成绩画像指标具体如表 2 所示。在表 2 中,学分指标体系拥有 1 个一级指标,二级指标按照课程类别、课程属性进行划分;成绩

Table 2 Introduction of student academic performance profile indicators

表 2 学生学业成绩画像指标介绍

指标体系	一级指标	二级指标
学分指标体系	总学分率	当前思政课学分率
		当前体育课学分率
		当前通识课学分率
		当前基础课学分率
		当前专业课学分率
		当前拓展课学分率
成绩指标体系	当前总绩点排名	当前通识课绩点排名
		当前基础课绩点排名
		当前专业课绩点排名
		当前拓展课绩点排名
		各课程绩点排名
	成绩波动程度	学期、学年成绩波动
		成绩波动方向
	总挂科率	通识课挂科率
		基础课挂科率
		专业课挂科率
		拓展课挂科率
		挂科课程总学分
课程指标体系	总课程优秀率	是否为优秀课程
		学期、学年课程优秀率
		通识课程优秀率
		专业课程优秀率
		基础课程优秀率
		拓展课程优秀率
		学期、学年优秀课程学分率
	总优秀课学分率	通识课优秀课程学分率
		专业课优秀课程学分率
		基础优秀课程学分率
		拓展优秀课程学分率
	低于课程均分率	学期、学年低于课程均分率
		通识课低于课程均分率
		专业课低于课程均分率
		基础优秀课程学分率
		拓展课优秀课程学分率
		学期、学年课程及格率
	总及格率	通识课程及格率
		专业课程及格率
		基础课程及格率
	偏科程度	拓展课程及格率
		偏科程度

指标体系拥有3个一级指标,当前总绩点排名二级指标按照课程类别进行划分,成绩波动程度二级指标按照学期学年时间线进行划分,总挂科率二级指标按照课程类别进行划分;课程指标体系拥有5个一级指标,共将学生课程分为3段,总优秀课程学分率是对总课程优秀率的补充,其次是低于课程均分率,最后为及格率,二级指标均按照课程种类或时间线进行划分。

本文使用分类算法的分类准确率来评估特征选择算法所选特征子集的优劣。因此,本文实验中对原始数据集与经过特征选择后的数据集使用C4.5决策树算法的分类准确率和最终选择特征的数量进行评估。本文实验包括基于UCI公共数据集实验和应用实验,之后再在学生画像指标数据集上进一步评估算法的应用能力。

4.1.2 实验设置

本文实验的机器配置参数如下:基于Intel® Core™ i56300HQ、2.6 GHz主频、16 GB内存以及Windows 10操作系统,实验仿真软件采用PyCharm, 2020.2版本。

参数设置会影响算法的全局收敛性能。控制参数实验被广泛用于调度优化、组合优化和函数优化等问题,具有易于理解、便于实现等优点[19,20]。因此,本文将控制参数实验用于算法参数的设定。通过实验设计,对粒子群优化算法的2个学习因子(c_1和c_2)进行设定。本文给出了参数选择表,如表3所示,共选取9组参数组合,并将式(10)作为适应度函数。由于算法的随机性等特点,本文将每组参数运行10次的结果取平均值作为最终适应度值。通过9组实验结果可以发现,学习因子c_1和c_2值为2时,算法的适应度值最低,算法的性能最好。为保证实验的公平性,最大迭代次数和种群规模均与对比算法的一致。

Table 3 Parameter selection table

表3 参数选择表

参数c_1	参数c_2	适应度值
2.0	1.0	5.79
2.0	1.5	5.86
2.0	2.0	5.17
1.5	1.0	6.84
1.5	1.5	6.87
1.5	2.0	7.12
1.0	1.0	7.35
1.0	1.5	7.14
1.0	2.0	7.89

因此,所用粒子群优化算法的参数设置如下:学习因子c_1和c_2值为2,粒子个数N值为30,最大迭代次数t_{max}值为100。

4.2 UCI公共数据集实验

4.2.1 UCI公共数据集实验数据集介绍

为了检验提出的RF-ATPSO算法的性能及稳定性,本文基于UCI公共数据集,将RF-ATPSO算法与传统特征选择算法(包括RELIEF-F、PSO、GWO、RFGWO和RFPSO算法)进行对比实验。

实验分别在6个UCI公共数据集上进行,通过计算各算法选出的特征子集的准确率来评估算法的性能。在每个数据集上取20次实验的实验结果,分别选取最优准确率(*Best*)和平均准确率(*Avg*)2个指标来度量不同算法的性能。表4展示了RF-ATPSO算法与传统特征选择算法在6个数据集上取得的分类准确率。

由表4可知,C4.5算法在其原始特征集合上的准确率均比经过特征选择后的准确率低,出现这种现象主要因为原始数据高维特征空间和特征高度冗余对C4.5的分类结果产生了较大影响,但是也存在经过特征选择后的特征子集辨识度变差的情况。

表5给出了RF-ATPSO算法与传统特征选择算法从6个数据集中提取的平均特征子集规模。由表5可知,基于RF-ATPSO算法对数据集进行特征选择后,特征空间维度明显减小。观察表4和表5可知,RF-ATPSO算法在Meu、Scadi、Cancer、Arrhythmia和HillValley 5个数据集上所选的特征子集规模最小且准确率最高,即能以最低的特征空间维度取得最高的准确率。总之,本文提出的RF-ATPSO算法在保证准确率的情况下,可以有效提高C4.5算法的运行效率。

进一步分析表4中的实验结果,可以发现:对比3种传统的Filter和Wrapper算法RELIEF-F、GWO、PSO可知,经过特征选择后,C4.5算法分类准确率均有不同程度的提高。2种Wrapper算法在不同数据集上的性能表现不同,在Meu、Scadi、Spambase和HillValley数据集上,PSO算法的结果最优,在Cancer和Arrhythmia数据集上,GWO算法的结果最优。整体而言,PSO算法要优于RELIEF-F和GWO算法,平均分类准确率较2种算法分别提高了7.68%和0.70%。在所选特征子集规模上,PSO算法在6个数据集上均优于GWO

Table 4 Classification accuracies achieved by RF-ATPSO algorithm and traditional feature selection algorithms on 6 datasets

表 4 RF-ATPSO 算法与传统特征选择算法在 6 个数据集上取得的分类准确率 %

数据集	准确率	原始特征集合	Filter	Wrapper		Hybrid		
			RELIEF-F	GWO	PSO	RFGWO	RFPSO	RF-ATPSO
Meu	*Best*	64.46	59.67	67.9	**69.83**	60.67	63.24	68.23
	Avg	63.23	56.30	66.33	66.93	59.25	61.57	**67.31**
Scadi	*Best*	86.36	86.36	90.91	93.96	90.91	95.26	**95.26**
	Avg	79.31	82.73	90.45	91.82	90.45	93.28	**95.00**
Spambase	*Best*	91.89	91.74	93.47	**94.20**	91.37	91.75	92.32
	Avg	90.60	90.30	92.01	**93.25**	90.29	90.86	91.46
Cancer	*Best*	68.67	71.43	77.11	78.31	79.51	79.51	**81.92**
	Avg	63.96	62.14	78.84	77.65	78.31	78.73	**81.68**
Arrhythmia	*Best*	74.42	76.01	85.27	82.17	84.49	84.49	**85.89**
	Avg	69.53	69.73	83.41	81.02	82.17	82.51	**83.45**
HillValley	*Best*	54.95	57.69	64.28	67.58	65.28	67.03	**70.33**
	Avg	50.66	51.48	63.29	62.94	63.53	64.49	**65.74**
平均		71.50	72.29	79.27	79.97	80.03	80.25	**81.54**

Table 5 Average sizes of feature subsets extracted by RF-ATPSO algorithm and traditional feature selection algorithms from 6 datasets

表 5 RF-ATPSO 算法与传统特征选择算法从 6 个数据集中提取的平均特征子集规模

数据集	原始特征集合	Filter	Wrapper		Hybrid		
		RELIEF-F	GWO	PSO	RFGWO	RFPSO	RF-ATPSO
Meu	71	56.80	36.05	34.70	29.50	**24.70**	30.55
Scadi	205	163.40	63.65	30.85	48.75	40.10	**9.30**
Spambase	57	45.60	29.75	25.20	25.70	21.50	**18.80**
Cancer	56	44.80	24.05	8.90	12.95	10.20	**4.75**
Arrhythmia	279	223.20	132.50	104.50	102.50	90.40	**70.50**
HillValley	100	80.00	36.85	36.95	36.95	31.80	**23.20**
平均	128	102.30	48.81	40.18	42.72	36.45	**26.20**

算法,平均特征子集规模比 GWO 算法的低 8.63。总体而言,PSO 的特征选择结果较 GWO 具有一定优势。

对比 3 种混合式算法可知,算法针对不同的数据集,性能可能也会有所区别。由表 4 可知,在 Scadi、Cancer、Arrhythmia 和 HillValley 数据集上,RF-ATPSO 平均分类准确率最高,较 RFGWO 和 RFPSO 算法的均有小幅度提升,分别为 1.51%,1.29%;在 Meu 数据集上,平均准确率最高,但其最高分类准确率表现并非最优;在所选特征子集规模上,RT-ATPSO 算法在除 Meu 外的 5 个数据集上,特征子集规模最小;对比本文提出的 RF-ATPSO 和其他特征选择算法可知,RF-ATPSO 算法在 Spambase 数据集上分类准确率未达到最优,但整体而言 RF-ATPSO 的平均分类准确率达到 81.54%,在所有数据集上均表现最优。

4.2.2 UCI 公共数据集收敛性对比

本节实验将 GWO、PSO、RFGWO、RFPSO 和 RF-ATPSO 算法进行对比分析,图 2 为 3 种封装式特征选择算法在 6 个数据集上的错误率收敛曲线。

从图 2 可以看出,在 Cancer、Arrhythmia 和 HillValley 数据集上,RF-ATPSO 算法的收敛曲线均在 GWO、PSO、RFGWO 和 RFPSO 算法的之下;在 Cancer 和 HillValley 数据集上,RF-ATPSO 算法拥有较低的初始适应度值,并且能快速收敛至全局最优解,在所有算法中收敛速度最快;在 Arrhythmia 数据集上,RF-ATPSO 算法在迭代前期收敛速度低于 RFPSO 和 RFGWO 算法的,但在第 30 次迭代时,可迅速跳出局部最优解,向全局最优解收敛;在 Scadi 数据集上,没有经初步特征选择的 PSO 算法收敛速度较慢,但其优于 GWO 和 RFGWO 算法,RF-ATPSO 算法初始和最终收敛值最低,具有较快的收敛速度;在 Meu 和 Spambase 数据集上,尽管 RF-ATPSO 算法没有取得最优的收敛效果,但 RF-ATPSO 算法的收敛曲线在 RFPSO 的之下,因此本文提出的改进策略有效,并

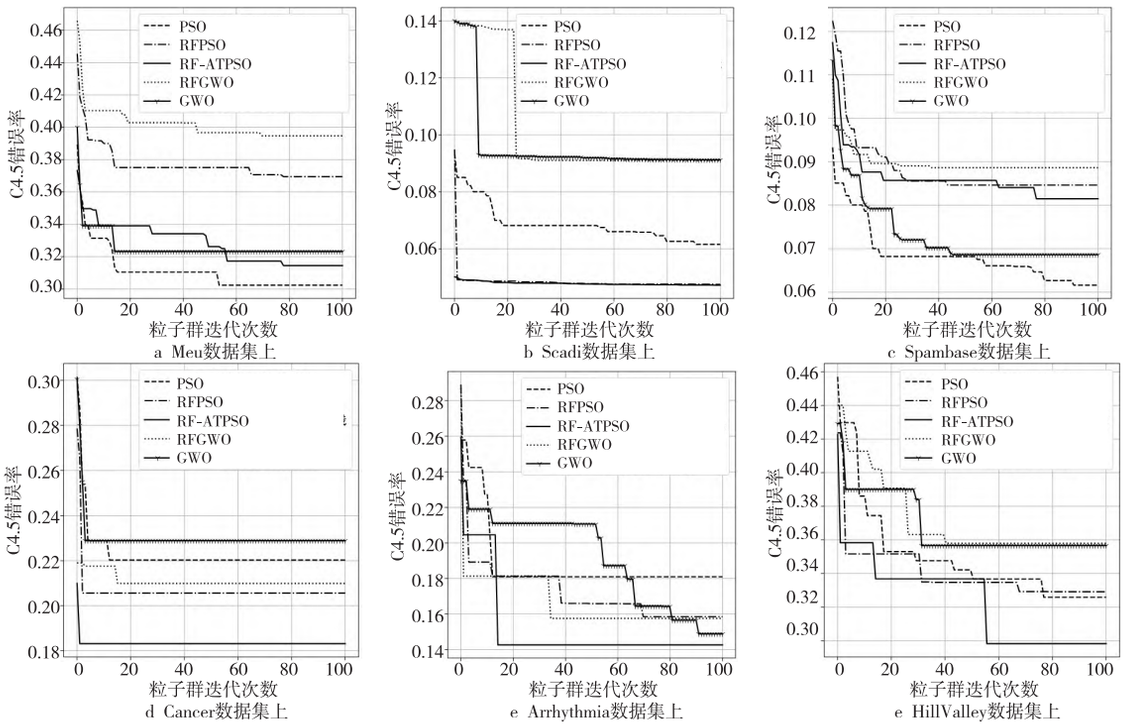

Figure 2 Error rate convergence curves

图 2 错误率收敛曲线

且利用 PSO 算法进行特征选择后均优于使用 GWO 算法的。经过 RELIEF-F 算法初步筛选特征后的 RFPSO 和 RF-ATPSO 算法收敛速度和收敛适应度值均不如 PSO 算法的,说明在上述 2 个数据集上 RELIEF-F 算法筛选过的特征子集本身辨识度差,在原特征空间中搜寻效果更佳。

4.3 学生学业成绩画像指标数据集实验

4.3.1 分类准确率和收敛性分析

为进一步验证 RF-ATPSO 算法的有效性,在表 2 所示的某高校学生学业成绩画像指标数据集上进行对比实验。选取计算机专业四年学业成绩数据,按照本文设计的特征指标体系,构建出的学生学业成绩画像拥有 227 维特征。实验中 RELIEF-F、GWO、PSO、RFGWO、RFPSO 和 RF-ATPSO 算法分别运行 20 次,分类准确率均值计算结果如表 6 所示。

由表 6 可以看出,C4.5 算法在原始数据集上的分类准确率较差,平均准确率仅为 88.51%,比用 RF-ATPSO 算法进行特征选择后的平均准确率低 6.26%。RF-ATPSO 算法在学生类别 1、2 及所有类别平均值上的准确率最高,尤其在类别 2 上准确率达到 94.82%,比原始数据集的准确率高 6.96%。

RF-ATPSO 算法相较于其他 5 种特征选择算法不仅总体准确率分别提高了 4.66%,2.64%,2.19%,2.15%和 1.98%,而且在 3 个类别上也均有不同程度的提高。在类别 1 和类别 2 上,RF-ATPSO 算法所求得的特征子集准确率最高,分别达到了 93.68%和 96.71%。

学生学业成绩画像指标数据收敛曲线如图 3 所示。从图 3 可知,RF-ATPSO 算法在收敛速度和收敛值方面,均优于其余4种特征选择算法;

Table 6 Classification accuracies of feature selection for the portrait index dataset

表 6 画像指标数据集特征选择的分类准确率 %

类别	原始特征集合	Filter	Wrapper		Hybrid		
		RELIEF-F	GWO	PSO	RFGWO	RFPSO	RF-ATPSO
类 1	85.67	86.92	88.91	89.15	92.62	93.02	**93.68**
类 2	89.75	92.96	92.48	93.74	93.14	92.76	**96.71**
类 3	91.18	90.63	93.88	**95.17**	93.68	94.28	94.82
平均	88.51	90.11	92.13	92.58	92.62	92.79	**94.77**

RF-ATPSO算法在第15次已寻找到全局最优解,证明其收敛速度较快,可及时跳出局部最优解;GWO、PSO分别在第50次和第19次寻找到全局最优值;RFGWO、RFPSO分别在第18次和第17次寻找到全局最优值。因此,RF-ATPSO不仅迭代次数少且最优解适应度值更低,拥有较高的寻优效率。

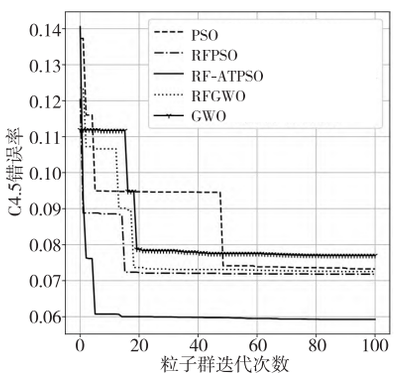

Figure 3 Convergence curves of student profile indicator data

图3 学生画像指标数据收敛曲线

4.3.2 RF-ATPSO特征选择结果分析

如前所述,本文构建的学生学业成绩画像经过RF-ATPSO算法特征选择后降到了82维,包括44个成绩指标(含8个课程排名指标、14个成绩波动指标、3个平均排名指标、18个绩点排名指标和1个挂科总学分指标);37个课程指标(含22个优秀课程指标、5个优秀课程学分率指标、6个低于均分课程率指标和4个优秀课程率指标)。

在实际数据中,学生出现挂科(不及格)的情况较少,因此,学分指标体系下各学生的各项指标值,大多接近于1,因此学分率的区分度不高,在特征选择结果中也基本没有学分率指标体系中的指标,可见该特征选择结果符合实际情况。在选择出的37个课程指标中,所有及格率相关指标的值均接近1,因此区分度不高,实际特征选择结果中,也没有及格率相关指标,可见该特征选择结果也符合实际情况。

5 结束语

针对高校教务领域数据固有的高维特征空间和高度冗余问题,本文提出了一种融合特征权重和改进粒子群优化算法的混合式特征选择算法(RF-ATPSO)。该算法主要分为2个步骤,首先使用RELIEF-F算法计算各个特征的权重,筛除冗余特征;然后从筛选出的特征集合中利用改进粒子群优化算法搜索最优特征子集。

实验方面,首先在6个UCI公共数据集上进行实验。结果表明,C4.5算法在经过RF-ATPSO算法特征筛选后的数据集上不仅准确率优于其他特征选择算法的,而且算法所选特征子集规模最小,在保证准确率的同时提高了C4.5算法的运行效率。在学生学业成绩画像指标数据集上的结果表明,C4.5算法在经过RF-ATPSO算法特征筛选后的数据集上准确率达到94.77%,优于其他传统特征选择算法。尽管本文提出的RF-ATPSO特征选择算法在大部分数据集上取得了较好效果,但还存在经RELIEF-F特征选择后特征子集辨识度变差的问题,未来将重点研究提高特征子集辨识度的最优方法。

参考文献:

[1] 刘譞. 基于学生行为的成绩预测模型的研究与应用[D]. 成都:电子科技大学,2018.
Liu Xuan. Research and application of performance prediction model based on student behavior[D]. Chengdu:University of Electronic Science and Technology of China,2018.

[2] Li J,Cheng K,Wang S,et al. Feature selection:A data perspective[J]. ACM Computing Surveys, 2017, 50(6): 94.1-94.45.

[3] 刘艺,曹建军,刁兴春, 等. 特征选择稳定性研究综述[J]. 软件学报,2018,29(9):2559-2579.
Liu Yi,Cao Jian-jun,Diao Xing-chun,et al. Survey on stability of feature selection [J]. Journal of Software,2018,29(9):2559-2579.

[4] 陈亮,汤显峰. 改进正余弦算法优化特征选择及数据分类[J]. 计算机应用,2022,42(6):1852-1861.
Chen Liang,Tang Xian-feng. Improved sine cosine algorithm for optimizing feature selection and data classification[J]. Journal of Computer Applications,2022,42(6):1852-1861.

[5] Sosa-Cabrera G,García-Torres M,Gómez-Guerrero S,et al. A multivariate approach to the symmetrical uncertainty measure:Application to feature selection problem[J]. Information Sciences,2019,494:1-20.

[6] Zhang X,Mei C,Chen D,et al. Feature selection in mixed data:A method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition,2016,56:1-15.

[7] 孙林,陈雨生,徐久成. 基于改进ReliefF的多标记特征选择算法[J]. 山东大学学报(理学版),2022,57(4):1-11.
Sun Lin,Chen Yu-sheng,Xu Jiu-cheng. Multilabel feature selection algorithm based on improved ReliefF[J]. Journal of Shandong University(Natural Science),2022,57(4):1-11.

[8] 李炜,巢秀琴. 改进的粒子群算法优化的特征选择方法[J]. 计算机科学与探索,2019,13(6):990-1004.
Li Wei,Chao Xiu-qin. Improved particle swarm optimization

method for feature selection[J]. Journal of Frontiers of Computer Science and Technology,2019,13(6):990-1004.

[9] Hu P,Pan J S,Chu S C. Improved binary grey wolf optimizer and its application for feature selection[J]. Knowledge-Based Systems,2020,195:105746.

[10] 吴晓燕,刘笃晋. 基于樽海鞘群与粒子群混合优化算法的特征选择[J]. 重庆邮电大学学报(自然科学版),2021,33(5):844-850.
Wu Xiao-yan,Liu Du-jin. Feature selection based on hybrid optimization of salp swarm algorithm and particle swarm optimization[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition),2021,33(5):844-850.

[11] Zhang Y,Wang S,Phillips P,et al. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection[J]. Knowledge-Based Systems,2014,64:22-31.

[12] Lu H,Chen J,Yan K,et al. A hybrid feature selection algorithm for gene expression data classification[J]. Neurocomputing,2017,256:56-62.

[13] 王金杰,李炜. 混合互信息和粒子群算法的多目标特征选择方法[J]. 计算机科学与探索,2020,14(1):83-95.
Wang Jin-jie,Li Wei. Multi-objective feature selection method based on hybrid MI and PSO algorithm[J]. Journal of Frontiers of Computer Science and Technology,2020,14(1):83-95.

[14] 肖艳,姜琦刚,王斌, 等. 基于 ReliefF 和 PSO 混合特征选择的面向对象土地利用分类[J]. 农业工程学报,2016,32(4):211-216.
Xiao Yan,Jiang Qi-gang,Wang Bin, et al. Object based land-use classification based on hybrid feature selection method of combining ReliefF and PSO[J]. Transactions of the Chinese Society of Agricultural Engineering,2016,32(4):211-216.

[15] 徐灯,傅晶,王文丰, 等. 一种加权变异的粒子群算法[J]. 南昌工程学院学报,2021,40(1):51-56.
Xu Deng,Fu Jing,Wang Wen-feng,et al. A weighted variation particle swarm optimization algorithm[J]. Journal of Nanchang Institute of Technology,2021,40(1):51-56.

[16] 杨鹤标,刘芳,胡惊涛. 基于 PSO 的小样本特征选择优化算法研究[J]. 江苏科技大学学报(自然科学版),2021,35(1):76-81.
Yang He-biao,Liu Fang,Hu Jing-tao. Research on one-shot learning feature selection algorithm based on particle swarm optimization[J]. Journal of Jiangsu University of Science and Technology(Natural Science Edition),2021,35(1):76-81.

[17] Kennedy J,Eberhart R. Particle swarm optimization[C]// Proc of International Conference on Neural Networks,1995:1942-1948.

[18] Kononenko I. Estimating attributes:Analysis and extensions of RELIEF[C]//Proc of the European Conference on Machine Learning,1994:171-182.

[19] 胡文斌,韩璞,孙明. 二进制编码遗传算法中的控制参数选取方法[J]. 计算机仿真,2015,32(3):447-450.
Hu Wen-bin,Han Pu,Sun Ming. Selection method of control parameters in binary coded genetic algorithm[J]. Computer Simulation,2015,32(3):447-450.

[20] Miret J,Balestrassi P P,Camacho A,et al. Optimal tuning of the control parameters of an inverter-based microgrid using the methodology of design of experiments[J]. IET Power Electronics,2020,13(16):3651-3660.

作者简介:

刘振超(1996 -),男,河北邢台人,硕士,CCF 会员(J0571G),研究方向为机器学习算法和群智能算法。**E-mail**:563950476@qq.com

LIU Zhen-chao, born in 1996, MS, CCF member(J0571G), his research interests include machine learning algorithm and swarm intelligent algorithm.

苑迎春(1970 -),女,河北保定人,博士,教授,博士生导师,CCF 会员(14279M),研究方向为智能信息处理和大数据。**E-mail**:nd_hd_yyc@163.com

YUAN Ying-chun, born in 1970, PhD, professor, PhD supervisor, CCF member(14279M), her research interests include intelligent information processing and big data.

王克俭(1971 -),女,河北泊头人,博士,教授,CCF 会员(19033M),研究方向为图文信息处理和智能算法。**E-mail**:wkj71@163.com

WANG Ke-jian, born in 1971, PhD, professor, CCF member(19033M), her research interests include image & text information processing and intelligent algorithm.

何晨(1996 -),男,河北张家口人,硕士,CCF 会员(J0665G),研究方向为自然语言处理、命名实体识别和知识图谱。**E-mail**:20202060092@pgs.hebau.edu.cn

HE Chen, born in 1996, MS, CCF member(J0665G), his research interests include natural language processing, named entity recognition, and knowledge graph.