文章编号:1007-5321(2021)04-0049-07　　**DOI**:10.13190/j.jbupt.2020-237

粒子群优化的模糊粗糙集双约简算法

刘占峰,　潘　甦
(南京邮电大学 江苏省通信与网络技术工程研究中心, 南京 210003)

摘要: 为了提升下游模型的性能,获得质量更好的约简数据集,提出基于粒子群优化(PSO)的模糊粗糙集特征和实例联合选择算法,引入基于 ε-双约简的适应度函数来评估约简集的质量,引导搜索过程快速逼近最优解. 实验结果表明,基于 PSO 算法的模糊粗糙集双约简算法有效约简了实例和特征,获得了高质量的约简集,在分类任务中取得了优于原始数据集的准确度.

关　键　词: 模糊粗糙集; 特征选择; 实例选择; 粒子群优化
中图分类号: TP181　　**文献标志码**: A

Fuzzy-Rough Bireducts Algorithm Based on Particle Swarm Optimization

LIU Zhan-feng,　PAN Su
(Jiangsu Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Selecting informative features and removing noise instances are beneficial to gain a clean dataset and promote the performance of subsequent classifiers. A novel algorithm for fuzzy-rough bireducts with particle swarm optimization is proposed. The fitness function with ε-bireduct is employed to evaluate the candidate fuzzy-rough bireducts, which drives the particle swarm optimization search process toward better candidate solutions. The selected optimal bireduct is utilized to construct the subsequent classifier. The experimental results show that the proposed algorithm is superior to the counterpart, which reduces the instances and features effectively, and obtains high-quality bireducts. The classification accuracy of the proposed algorithm is thus better than the counterpart.

Key words: fuzzy-rough; feature selection; instance selection; particle swarm optimization

特征选择[1-2]和实例选择[3]是 2 种常见的用于数据清洗和约简的方法. 将 2 种方法结合起来对数据集进行特征选择和实例选择(以下简称为双约简),可以找到高质量的约简数据集. 这里讨论模糊粗糙集方法中的双约简[4]问题.

作为粗糙集理论[5]一种广泛应用的扩展形式,模糊粗糙集可以有效地处理实际场景中的不确定和含糊性问题,被成功应用于数据预处理领域[6]. 同时模糊粗糙实例和特征选择(SFRIFS, simultaneous fuzzy-rough instance and feature selection)算法[7]是典型的基于模糊粗糙集的双约简算法,它本质上是一种贪婪解决方案,因此获得的约简集不太可能是

收稿日期: 2020-11-23
基金项目: 江苏省研究生科研与实践创新计划项目(KYCX18_0882);国家自然科学基金项目(6201244);江苏省重点研发计划项目(BE2018733)
作者简介: 刘占峰(1980—), 男, 博士生, E-mail: zf.liu@139.com; 潘　甦(1969—), 男, 教授, 博士生导师.

最优解. 为了改进 SFRIFS 算法的贪婪选择策略，Parthalain 等[8-9]提出基于和声搜索的启发式策略寻找模糊粗糙双约简集，简称为 HSFSBR（harmony search based feature selection algorithm fuzzy-rough bi-reducts）算法，但和声搜索的随机探索机制生成高质量的新和声概率较低，和声记忆库在迭代过程中更新较慢[10].

为了弥补已有模糊粗糙集双约简算法的不足，提出了一种基于粒子群优化（PSO, particle swarm optimization）[11]算法的模糊粗糙集双约简算法，根据粒子自身和邻居粒子探索遇到的最优特征组合共同确定模糊粗糙集双约简集[12]. PSO 算法在复杂搜索空间中具有健壮的探索能力[13]，被广泛应用在人脸识别[14]、金融风控[15]和特征工程[16]等领域. 笔者的主要贡献是在基于 PSO 算法的粗糙集特征选择方法基础上实现了模糊粗糙集双约简算法，综合考虑特征和实例数量，寻找质量更好的双约简集. 与已有的模糊粗糙集双约简算法相比，所提算法的更新机制避免了贪婪搜索和随机特征选择，发现了更优的约简集. 实验结果表明，在相同的实验条件下，所提算法显著减少了特征和实例数量，同时又保持了更高的分类精度.

1 背景知识

在粗糙集理论中，信息系统表示为$\langle U,A\rangle$，其中$U=\{x_1,x_2,\cdots,x_{|U|}\}$和$A=\{a_1,a_2,\cdots,a_{|A|}\}$分别为实例和特征的有限非空集合，$A$中的每一项$a$都对应$U\to V_a$的映射，其中$V_a$是$a$在实例集合$U$上的值域. 通过包含决策特征可以把信息系统扩展为决策系统$\langle U,A\cup\{d\}\rangle$，其中$d(d\notin A)$为决策特征.

1.1 模糊粗糙集

粗糙集理论仅适用于离散数据，而实际场景中的特征值往往是连续值. 为了进一步有效处理连续数据，引入模糊粗糙集[17]理论.

模糊粗糙集把离散情况下的上下近似公式扩展到连续数据领域. 在离散情况下，实例要么属于下近似，要么不属于下近似. 而在连续情况下，实例用取值在[0,1]之间的隶属度来表示对不确定性的刻画，离散情况相当于隶属度取值为0或1的极端情况. 根据定义上下近似公式，模糊粗糙集特征选择[18]将模糊粗糙集的思想应用于特征选择，有

$$\mu_{\underline{R_B}X}(x_i)=\inf_{x_j\in U}I(\mu_{R_B}(x_i,x_j),\mu_X(x_j)) \tag{1}$$

$$\mu_{\overline{R_B}X}(x_i)=\sup_{x_j\in U}\mathcal{T}(\mu_{R_B}(x_i,x_j),\mu_X(x_j)) \tag{2}$$

其中：X为模糊集，I为模糊算子，$\mathcal{T}$为t-norm算子，R_B为对于特征子集B的模糊相似关系. 对于任意的$x_i,x_j\in X$，有

$$\mu_{R_B}(x_i,x_j)=\mathcal{T}_{a\in B}\{\mu_{R_a}(x_i,x_j)\} \tag{3}$$

其中$\mu_{R_a}(x_i,x_j)$是对特征$a(a\in A)$而言实例x_i和x_j之间的相似度. 常用的3种相似度的关系定义为（以下简称为sim1～sim3）：

sim1：

$$\mu_{R_a}(x_i,x_j)=1-\frac{|a(x_i)-a(x_j)|}{a_{\max}-a_{\min}} \tag{4}$$

sim2：

$$\mu_{R_a}(x_i,x_j)=\exp\left(-\frac{[a(x_i)-a(x_j)]^2}{2\sigma_a^2}\right) \tag{5}$$

sim3：

$$\mu_{R_a}(x_i,x_j)=\max\left(\min\left(\frac{a(x_j)-[a(x_i)-\sigma_a]}{\sigma_a},\frac{[a(x_i)+\sigma_a]-a(x_j)}{\sigma_a}\right),0\right) \tag{6}$$

其中：σ_a为特征a的方差，$a(x_i)$为实例x_i在特征a上的取值.

1.2 模糊分辨矩阵

作为离散分辨矩阵的扩展，模糊分辨矩阵用模糊短语C表示. 模糊分辨矩阵的每一项都是模糊集，假设每个特征$a\in A$属于该项的概率为$\mu_{C_{ij}}(a)$，计算公式为

$$\mu_{C_{ij}}(a)=\mathcal{N}(\mu_{R_a}(x_i,y_j)) \tag{7}$$

其中$\mathcal{N}$为模糊否定算子.

模糊短语用于构造模糊分辨函数为

$$f_C(B)=f_C(a_1^*,a_2^*,\cdots,a_{|A|}^*)=\wedge\{\vee C_{ij}^*\mid C_{ij}\in C\} \tag{8}$$

其中$1\leqslant i\leqslant j\leqslant|U|$，$a_i^*$是给定特征$a_i\in A$的析取范式.

$$a_i^*=\begin{cases}\text{真}, & \text{当 } a_i\in B\\ \text{假}, & \text{其他}\end{cases} \tag{9}$$

函数返回值在0～1之间，反映了将指定变量$\{a_1,a_2,\cdots,a_{|A|}\}$分配真值时函数在多大程度上被满足.

对决策系统考虑不同决策值的短语，按照下列逻辑运算修正模糊分辨函数：

$$f_C(a_1^*,a_2^*,\cdots,a_{|A|}^*)=\wedge\{\vee\{C_{ij}^*\}\to\neg\ (d(x_i)=d(x_j))\} \tag{10}$$

对特定的特征子集 B,短语 C_{ij}的满足度定义为

$$S_B(C_{ij}) = S_{a\in B}\{\mu_{C_{ij}}(a)\} \tag{11}$$

为了适用于双约简场景,修正模糊分辨函数[7]使其包含实例因素:

$$f_C(B,Y) = f_C(a_1^*,a_1^*,\cdots,a_{|A|}^*,x_1^*,x_2^*,\cdots,x_{|U|}^*) = \wedge\{x_i^* \vee x_j^* \vee C_{ij}^*\}$$

$$x_i^* = \begin{cases} 真, & 当\ x_i \in Y \\ 假, & 其他 \end{cases} \tag{12}$$

此时,如果 C_{ij}^* 被最大程度地满足,则短语 C_{ij}被特征满足;如果选择了 x_i 或 x_j,则短语 C_{ij}被实例满足. 实例被选择意味着相应的训练实例将被删除,这些被选择的实例构成了离群点 $O = U/Y$.

2　基于 PSO 算法的模糊粗糙双约简算法

2.1　PSO 算法

PSO 算法是一种群体智能算法,初始化为一群随机粒子,通过迭代找到最优解. 每次迭代中,粒子跟踪自身的局部最优位置 p_b 和群体全局最优位置 g_b 来更新自己. 在得到这 2 个最优位置后,粒子通过式(13)和式(14)更新自己的速度和位置.

$$V_i = V_i + c_1\text{rand}()(p_{b,i} - x_i) + c_2\text{rand}()(g_{b,i} - x_i) \tag{13}$$

$$x_i = x_i + V_i \tag{14}$$

其中:$i = 1,2,\cdots,N$,N 为该群体中粒子的总数;V_i 为粒子的速度;rand( ) 为介于(0,1)之间的随机数;x_i 为粒子的当前位置. c_1 和 c_2 为学习因子. 式(13)和式(14)构成了 PSO 算法的标准形式,但标准 PSO 算法容易陷入局部最优. 此处采用修正的 PSO 算法,引入惯性加权因子,速度更新公式变为

$$V_i = \omega V_i + c_1\text{rand}()(p_{b,i} - x_i) + c_2\text{rand}()(g_{b,i} - x_i) \tag{15}$$

其中:ω 为惯性加权因子,由文献[13]得到,有

$$\omega = (\omega - 0.4)\frac{(i_{\max} - i)}{i_{\max}} + 0.4 \tag{16}$$

其中:$i_{\max}$为最大迭代次数,i 为当前迭代次数.

2.2　基于 PSO 算法的粗糙集特征选择

在基于 PSO 算法的粗糙集特征选择算法中,粒子映射为特征子集,位置用长度为 N 的二进制字符串表示,这里表示特征总数. 每位代表一个特征,1 表示选中对应位置的特征,0 则为未选中.

粒子的速度取 1 与最大速度之间的一个正整数,决定每次迭代改变多少个特征. 粒子位置间不相同位的数量代表位置差异程度. 最大速度控制粒子的全局探索能力. 当最大速度过小时,粒子很难逃离局部最优区域的束缚;但如果速度过大,粒子可能会飞过最优位置.

基于 PSO 算法的粗糙集特征选择算法只适用于特征选择任务,笔者针对双约简应用场景提出了 PSOBR(PSO based bireduct)算法.

2.3　PSOBR 算法

PSOBR 算法引入 ε-双约简的适应度函数,将其扩展到支持特征和实例联合选择领域. 适应度函数是一类特殊的目标函数,由用户定义评估优化效果的指标. 与深度学习中的损失函数类似,针对特定的优化目标构造具体的适应度函数表达式,引导搜索过程向更好的候选解方向运动. 双约简集(B,Y)的最优性可以从不同的角度评估,例如,可从分类质量和特征子集的长度来评估双约简集质量[13]. 在本算法中,考虑特征子集$|B|$的大小和选择实例$|Y|$的多少,引入 ε-双约简[19]以评估双约简集的质量.

不同的 ε 值直接影响 Y 的大小,从而影响特征子集的大小. 对特定的 ε,为了找到最小的双约简集,粒子群 s(相应的特征子集为 B_s)的适应度计算公式定义为

$$D_f(s) = \begin{cases} \text{cov}(B_s), & \text{cov}(B_s) \leqslant 1-\varepsilon \\ 2-2\varepsilon-\text{cov}(B_s), & \text{cov}(B_s) > 1-\varepsilon \end{cases} \tag{17}$$

此处 $\text{cov}(B_s)$表示根据特征子集 B_s 可分辨实例的最大比例:

$$\text{cov}(B_s) = \max_{Y\in \boldsymbol{Y}_{B_s}}\left(\frac{|Y|}{|U|}\right) \tag{18}$$

从适应度函数公式可以看出,最优适应度由 ε 决定. 对于特定的 ε 值,适应度函数调整 $\text{cov}(B_s)$直至达到最大适应度 $1-\varepsilon$.

所提方法用于模糊粗糙集特征和实例联合选择,算法流程如下.

1) 初始化搜索空间. 初始化参数,包括粒子群大小、学习因子 c_1 和 c_2、惯性加权因子 ω、粒子的最大速度、大迭代次数和最大适应度等超参数. 随机生成粒子的位置和速度,形成搜索空间.

2) 探索搜索空间. 遍历计算每个粒子即特征子集的适应度,与初始的自身 p_b 和 g_b 比较,用适应度更高的粒子更新 p_b 和群体的 g_b.

3) 更新搜索空间. 根据上一步的探索结果,遍

历每个粒子,根据当前的 p_b 和 g_b,用式(15)和式(14)更新速度和位置,完成搜索空间的更新.

4) 迭代. 重复步骤2)和步骤3),直到取得最大迭代次数或最大适应度. 适应度最大的双约简集 (B,Y) 则为最终的搜索结果.

3 实验结果

为了证明PSOBR算法的有效性,从UCI机器学习数据集[20]中选择8个数据集与其他模糊粗糙双约简算法[7-9]进行性能比较. 初始化参数如表1所示. 粒子速度控制粒子的全局探索能力,将其设置在 $[1,(1/3)N]$[13]. 惯性加权因子如式(16)定义,常数 c_1 和 c_2 取默认值[20]. 由于PSO算法的性能对粒子群的大小不敏感[20],取典型值为20.

与其他双约简算法一样,分别采用J48,JRip,PART和VQNN四种分类器做10折交叉来验证约简集的质量. 实验设置固定的随机种子,以保证实验结果可复现.

表1 参数设置及数据集信息

粒子群大小	最大迭代次数	c_1	c_2	惯性加权因子	粒子速度	最大适应度
20	20	2.0	2.0	1.4~0.4	$1\sim(1/3)N$	1

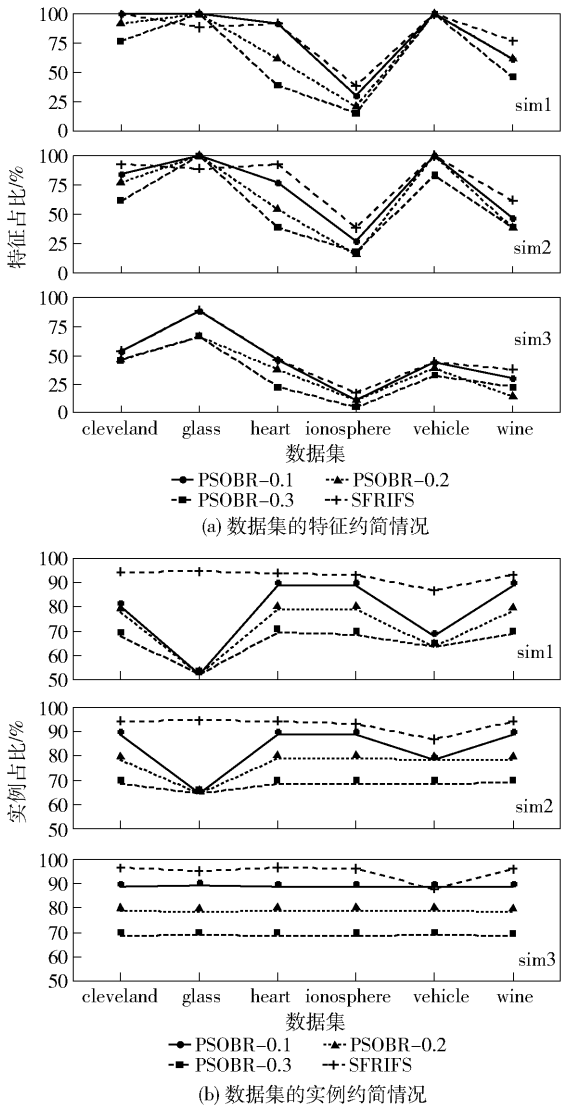
3.1 PSOBR算法和SFRIFS算法的比较

PSOBR算法和SFRIFS算法对数据集约简后特征和实例占原数据集的比例情况如图1所示. 在PSOBR算法中,大部分情况下约简后实例的比例接近于相应 ε 约束下的最优值. 与原始数据集相比,2种算法选择的特征都有不同程度地减少. 但是,从实验结果来看,PSOBR算法在获得更精简的约简集上能力更强.

2种算法在约简集上的分类精度比较结果如图2~4所示. SFRIFS算法下分类精度最高的wine数据集在PSOBR算法下仍然保持了相近的精度,但需要的实例和特征却少了很多. SFRIFS算法下分类不理想的数据集在PSOBR算法下获得了明显的性能提升. 例如,在SFRIFS算法下分类效果最差的cleveland数据集获得了将近20%的性能提升. 分析结果表明,PSOBR算法获得了更精简的高质量约简集.

3.2 PSOBR算法和HSFSBR算法的比较

图5所示为2种算法选择特征比例的对比情况. 可以看出,当采用sim3时,2种算法选择的特征



(a) 数据集的特征约简情况

(b) 数据集的实例约简情况

图1 PSOBR和SFRIFS算法的约简结果

数量相当. 但是,另外2种相似度关系在约简集中保留了更多的特征. 从图6的结果可以看出,在大多数情况下应用PSOBR算法的分类精度优于HSFSBR算法. 从图7可见,随着 ε 的增大,分类精度的提升效果明显. 采用sim1和sim2时可获得令人满意的分类能力. 在相同 ε-双约简约束下,PSOBR算法保留了更多的特征和原数据集性能,总体上获得了优于HSFSBR算法的性能.

3.3 适应度收敛分析

图8所示为改变粒子最大速度时,sonar数据集的适应度随迭代次数的变化趋势. 当迭代次数增加时,适应度快速收敛到特定 ε 对应的最优值. 当粒子的最大速度取值过小或过大时,由于陷入局部最优或飞过最优区域,在探索初期会出现一段平坦区

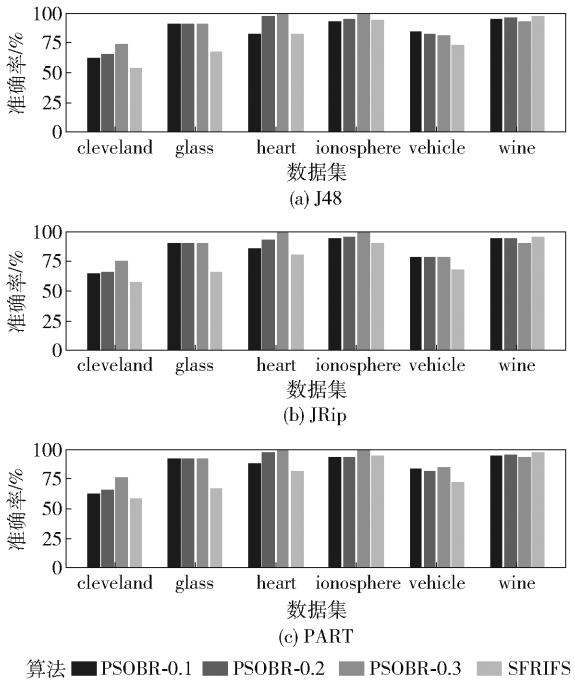


图 2　PSOBR 和 SFRIFS 算法在 sim1 时的分类结果

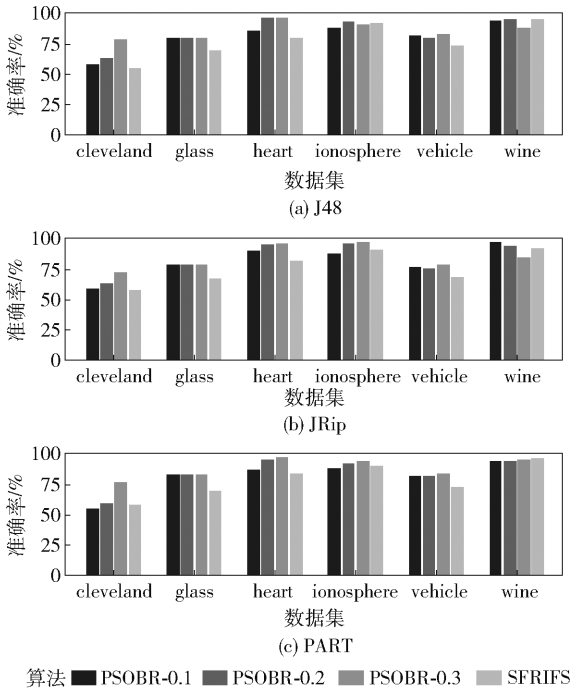


图 3　PSOBR 和 SFRIFS 算法在 sim2 时的分类结果

域,但最终都收敛到最优值附近.

迭代次数增加到 1 000 次时适应度的收敛情况如图 9 所示. 可以看出,适应度快速收敛到最优值,与 HSFSBR 算法设定的迭代次数相比, PSOBR 算法在问题空间中用相对较少的迭代次数找到最小约简集.

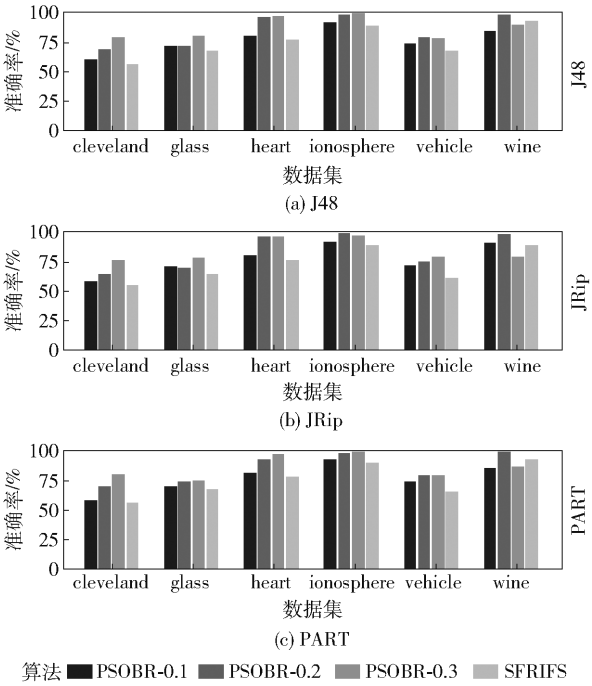


图 4　PSOBR 和 SFRIFS 算法在 sim3 时的分类结果

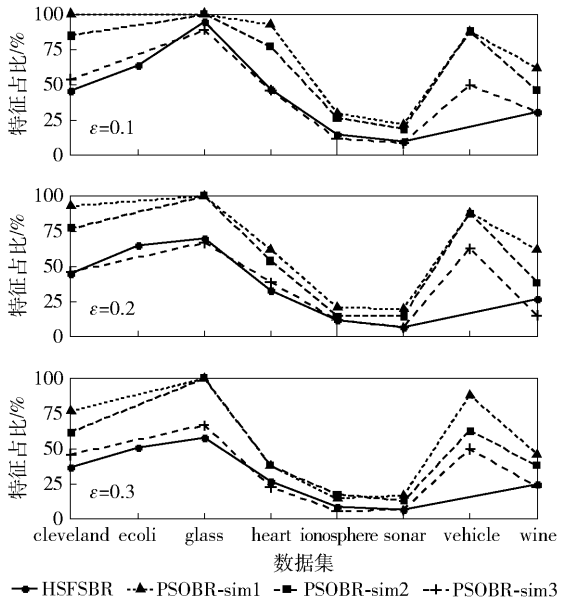
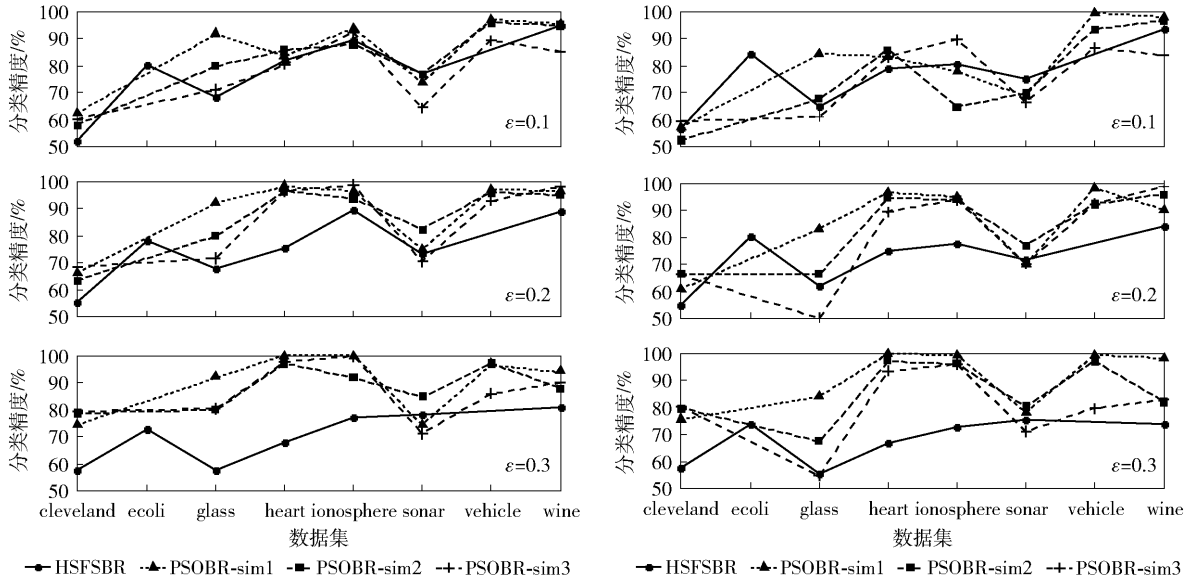


图 5　PSOBR 和 HSFSBR 约简集的特征数比较

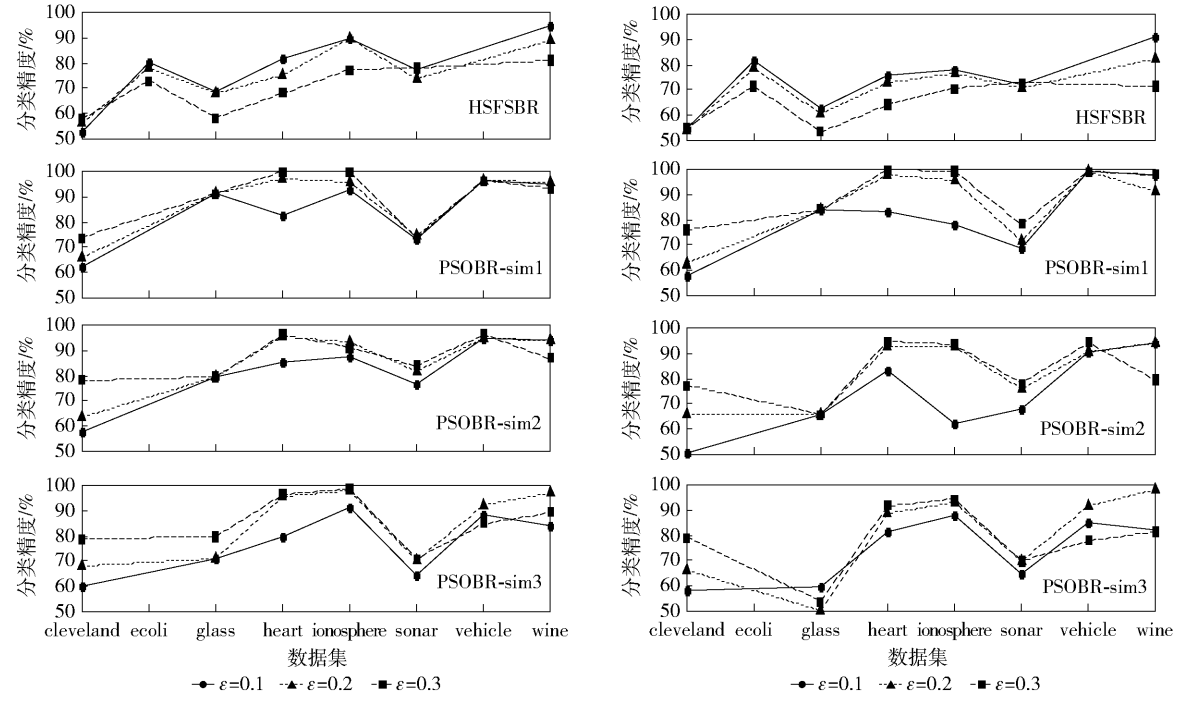
4　结束语

提出了一种新的算法用于在特征和实例联合选择任务中识别高质量的模糊粗糙双约简集. 为了克服同类算法在最优解搜索过程中的不足,引入 PSO 算法指导搜索过程. 此外用 ε-双约简的概念实现了特征和实例数量的折中,以找到合适的候选解. 提出的算法在有限的迭代次数内获得了高质量的双约



(a) C4.5分类器下的分类精度 (b) VQNN分类器下的分类精度

图6 HSFSBR与PSOBR分类精度的比较1



(a) C4.5分类器下的分类精度 (b) VQNN分类器下的分类精度

图7 HSFSBR与PSOBR分类精度的比较2

简集,在更精简的数据集上获得了更好的分类精度.

参考文献:

[1] Hancer E, Xue B, Zhang M. A survey on feature selection approaches for clustering[J]. Artificial Intelligence Review, 2020, 53(2): 4519-4545.

[2] 张东方, 陈海燕, 王建东. 半监督特征选择综述[J/OL]. 计算机应用研究, 2021, 38(2)[2020-10-13]. http://www.arocmag.com/article/02-2021-02-003.html.

[3] Derrac J, García S, Herrera F, et al. A survey on evolutionary instance selection and generation [J]. International Journal of Applied Metaheuristic Computing, 2017, 1(1): 60-92.

[4] Ślęzak D, Janusz A. Ensembles of bireducts: towards robust classification and simple representation [C] //

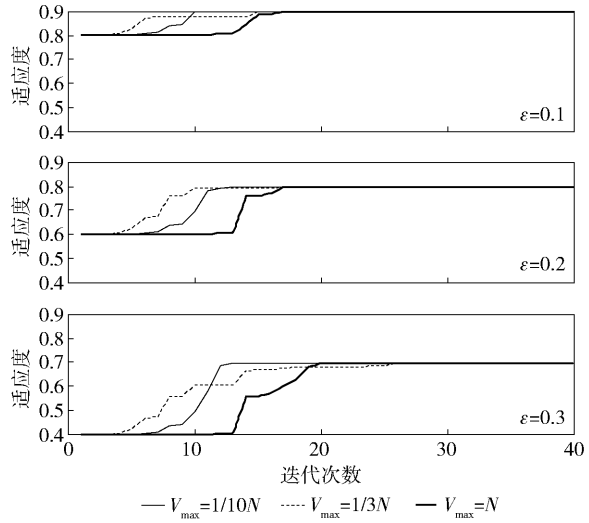


图 8　最大速度对粒子探索收敛的影响

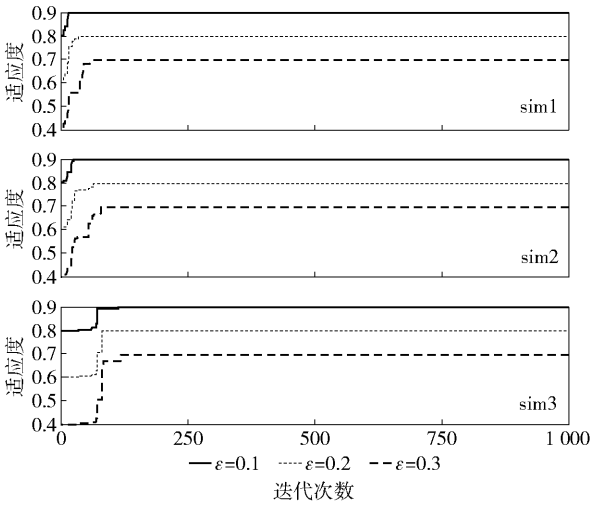


图 9　适应度曲线

International Conference on Future Generation Information Technology. Berlin: Springer, 2011: 64-77.

[5] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. [S. l.]: Kluwer Academic Publishers, 1992.

[6] Jensen R, Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches[J]. Kybernetes, 2008, 38(3/4): 438.

[7] Mac Parthalain N, Jensen R. Simultaneous feature and instance selection using fuzzy-rough bireducts[C]// IEEE International Conference on Fuzzy Systems. [S. l.]: IEEE, 2013: 1-8.

[8] Mac Parthalain N, Jensen R, Diao R. Fuzzy-rough set bireducts for data reduction[J]. IEEE Transactions on Fuzzy Systems, 2019 (99): 1-1.

[9] Diao R, Mac Parthalain N, Jensen R, et al. Heuristic search for fuzzy-rough bireducts and its use in classifier ensembles[C]//2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). [S. l.]: IEEE, 2014: 1504-1511.

[10] Ammar M, Bouaziz S, Alimi A M, et al. Hybrid harmony search algorithm for global optimization[C]//Nature and Biologically Inspired Computing, 2013 World Congress on. [S. l.]: IEEE, 2013.

[11] Kennedy J, Eberhart R. Particle swarm optimization[C]//Proceedings of ICNN'95- International Conference on Neural Networks. [S. l.]: IEEE, 1995: 1942-1948.

[12] Tu C J, Chuang L Y, Chang J Y, et al. Feature selection using PSO-SVM[J]. IAENG International Journal of Computer Science, 2007, 33(1): 111-116.

[13] Wang X, Yang J, Teng X, et al. Feature selection based on rough sets and particle swarm optimization[J]. Pattern Recognition Letters, 2007, 28(4): 459-471.

[14] Mistry K, Zhang L, Neoh S C, et al. A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition[J]. IEEE Transactions on Cybernetics, 2016, 47(6): 1496-1509.

[15] 孙艺, 宋振铭, 赵佳琪. 粒子群算法在金融风险模型中的研究与改进[J]. 吉林大学学报(信息科学版), 2020, 38(2): 85-91.
Sun Yi, Song Zhenming, Zhao Jiaqi. Research and improvement of particle swarm optimization in financial risk model[J]. Journal of Jilin University (Information Science Edition), 2020, 38(2): 85-91.

[16] Tran B, Xue B, Zhang M. A new representation in PSO for discretization-based feature selection[J]. IEEE Transactions on Cybernetics, 2017, 48(6): 1733-1746.

[17] Dubois D, Prade H. Putting rough sets and fuzzy sets together[M]//Intelligent Decision Support. Dordrecht: Springer, 1992: 203-232.

[18] Jensen R, Shen Q. New approaches to fuzzy-rough feature selection[J]. IEEE Transactions on fuzzy systems, 2008, 17(4): 824-838.

[19] Stawicki S, Ślęzak D. Recent advances in decision bireducts: complexity, heuristics and streams[C]//International Conference on Rough Sets and Knowledge Technology. Heidelberg: Springer, 2013: 200-212.

[20] Shi Y, Eberhart R C. Empirical study of particle swarm optimization[C]//Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406). [S. l.]: IEEE, 1999: 1945-1950.