doi:10.11835/j.issn.1000-582X.2020.01.008

基于多目标蜂群算法的数据分类方法

王海泉[1a],侯宇亮[1b],魏建华[1b],徐晓滨[2],苏孟豪[1b],张姗姗[1b]
(1.中原工学院 a.中原彼得堡航空学院;b.电子信息学院,郑州 450007;2.杭州电子科技大学 自动化学院, 杭州 310018)

摘要:为了保证运算时效的同时,提高复杂数据的分类精度,提出了基于多目标蜂群算法和极限学习机的数据分类算法。该方法以最小的特征个数和最高的分类精度为优化目标,利用改进的多目标蜂群算法对数据的特征个数和分类器参数进行寻优,针对多个有代表性的数据集进行仿真,结果表明所提出方法的有效性。

关键词:蜂群算法;多目标优化;特征选择;极限学习机;数据分类
中图分类号:TP181 **文献标志码**:A **文章编号**:1000-582X(2020)01-074-08

Research of data classification method based on multi-objective artificial bee colony algorithm

WANG Haiquan[1a], *HOU Yuliang*[1b], *WEI Jianhua*[1b], *XU Xiaobin*[2], *SU Menghao*[1b], *ZHANG Shanshan*[1b]
(1a. Zhongyuan Petersburg Aviation College; 1b. School of Electronic Information, Zhongyuan University of Technology, Zhengzhou 450007, P. R. China; 2.School of Automation, Hangzhou Dianzi University, Hangzhou 310018, P. R. China)

Abstract: In order to improve the classification accuracy of complex data on the premise of ensuring operation efficiency, a data classification algorithm based on multi-objective artificial bee colony algorithm and extreme learning machine is proposed, it takes the number of features and the classification accuracy as the optimization objectives, and improved artificial bee colony algorithm is introduced to optimize the parameters of the classifier and the selection of features of data. The simulation results based on six data sets verify the effectiveness of the proposed method.

Keywords: artificial bee colony algorithm; multi-objective optimization; feature selection; extreme learning machine; data classification

随着互联网技术的发展,文本数据在维度和样本数量上都出现了指数增长,对多维数据进行高效且有效的分析变得越来越具有挑战性,而基于机器学习的分类技术以其能够从海量数据中自动发现模式和知识的能力得到广泛应用。这其中极限学习机(Extreme learning machine,ELM)方法作为一种新型高效的单隐含层前馈神经

收稿日期:2019-05-13
基金项目:河南省高等学校重点科研项目(18A120005);浙江省重点研发项目(2019C03104)。
Support by Program of Educational committee of Henan province (18A120005) and Zhejiang Province Key R&D projects (2019C03104).
作者简介:王海泉(1981—),男,博士,副教授,主要从事数据挖掘、非线性控制等方向研究,(E-mail) wanghq@zut.edu.cn。

网络学习方法，克服了传统的基于梯度下降法的 BP 网络参数多、速度慢的缺点，已成为学者们研究的热点。

为了达到多维数据的最佳分类效果，如何选择最少最有效的特征和最优的 ELM 分类器参数成为分类器设计的关键。文献[1]使用粒子群算法优化隐含层节点数，文献[2]提出了基于遗传算法(GA)的 KELM 分类器参数优化方法，文献[3]利用改进的人工蜂群算法优化极限学习机的参数，有效改善了分类模型的效果。文献[4]将极限学习机与粒子群算法相结合，用于优化极限学习机相关参数，并将该模型应用于山洪的预测。文献[5]提出了一种基于遗传算法的 wrapper 特征选择方法，并利用极限学习机，建立了 ICU 并行死亡率预测模型。文献[6]将改进粒子群算法 MPSO 和极限学习机相结合，利用 MPSO 对单隐层前馈神经网络的隐节点参数进行优化。文献[7]采用灰太狼优化智能算法优化极限学习机相关参数，构建了一个有效的 KELM 模型。文献[8]提出一种改进粒子群优化算法，并将该算法用于优化极限学习机的惩罚系数和核宽，以提高模型预测精度和泛化性能。文献[9]采用改进的花朵授粉算法优化 ELM 预测模型的初始参数，提高了准确率和运行效率。可以看出，对于特征个数的选择和 ELM 分类器参数的优化问题，相关研究已经非常深入，但同时实现特征和 ELM 分类器参数选择的研究并不多见。为了获得更优的分类效果，文中结合参数少、鲁棒性强的人工蜂群算法，同时完成高维数据的特征选择和 ELM 分类器参数的寻优。

1　极限学习机原理

极限学习机是由 Huang 等[10]于 2004 年提出的前馈、单隐含层快速神经网络算法。

对于一个含有 N 个不同的训练样本集$\{(x_i,y_i)\mid x_i\in R^n,y_i\in R^m\}$，其隐含层网络的激活函数为 $g(x)$，ELM 分类模型可表示为

$$f=\sum_{f=1}^{L}\beta_j G(W_j\cdot X_i+b_j)=Y_i,i=1,2,...,N, \tag{1}$$

其中，L 为隐层节点数；$G(x)$ 为激励函数；$W_j=[w_{j1},w_{j2},\cdots w_{jm}]^{\mathrm{T}}$ 为第 j 个隐层节点与输入节点的权重向量；$\beta_j=[\beta_{j1},\beta_{j2},\cdots\beta_{jn}]^{\mathrm{T}}$ 为第 j 个隐层节点与输出节点的权重向量；b_j 为第 j 个隐层节点的偏置。ELM 训练步骤如下：

1)设定隐含层网络的节点数，随机设定输入层网络和隐含层网络的连接权重 W 和隐含层网络节点的阈值 $\boldsymbol{b}$；

2)选择一个无限可微的函数，作为隐含层网络的激活函数，并计算隐含层输出矩阵 $\boldsymbol{H}$；

3)计算输出层权值 $\boldsymbol{\beta}$。

显然在训练分类器过程中，隐含层和输出层之间的连接权值通过解方程组方式一次性确定，而隐含层节点数 L 需要设置，输入层和隐含层的全连接权值 W、隐含层激活函数阈值 $\boldsymbol{b}$ 等参数可以随机生成或手动调整。为了保证最优分类效果，文中引入人工蜂群算法，同时对 W、b 和数据特征数联合寻优，从而得到最优分类器。

2　基于蜂群算法的分类器优化

人工蜂群算法(artificial bee colony algorithm，ABC)是由土耳其学者 Karaboga 于 2005 年提出的一种新型群智能优化算法。该方法搜寻最优解的过程可分为 3 个阶段。第一是采蜜蜂阶段：采蜜蜂执行邻域搜索，并基于贪婪准则评估适应度值；之后是观察蜂阶段，它依据采蜜蜂阶段获得的解的质量好坏，执行同采蜜蜂相同的搜索；最后是侦察蜂阶段，当解更新次数达到局部搜索上限时，蜂的角色发生转换，放弃旧解，采用随机全局搜索方式生成一个新解。3 个阶段搜索完成后，最终保留寻找到的最优解。

2.1　解的编码及更新

在利用 ABC 算法进行分类器优化时，解向量由 ELM 分类器的参数 W、$\boldsymbol{b}$ 和特征向量 $\boldsymbol{f}$ 三部分组成，且均为[−1,1]间的实数编码，解的编码形式如下：

$$\begin{aligned}x=[&w_{11},w_{12},\cdots w_{1n},w_{21},w_{22},\cdots w_{2n},\cdots,w_{l2},\\&\cdots w_{ln},b_1,b_2,\cdots b_l,f_1,f_2\cdots f_n],\end{aligned} \tag{2}$$

其中，n、l 分别为输入网络、隐含网络节点数。

在解的更新过程，从 W、$\boldsymbol{b}$、$\boldsymbol{f}$ 三部分中分别随机选择一维同时进行更新，如式(3)所示，式中 α、β 为区间

[−1,1]之间的随机数。

$$v_{ij} = x_{ij} + \alpha(x_{ij} - x_{kj}) + \beta(\mathrm{Rpop}_{ij} - x_{ij}), \tag{3}$$

其中:,$k \neq i$,V_{ij}为更新后第 i 个解的第 j 维,x_{ij}为第 i 个解的第 j 维,Rpop_{ij}为帕累托解集中第 i 个解的第 j 维。

2.2 适应度函数构造

适应度函数用于衡量待优化解的质量,也就是分类器的性能。文中是以数据特征个数最少、分类器精度最高这样 2 个相互矛盾的指标为目标,如式(4)所示:

$$\begin{cases} f_1 = N, \\ f_2 = \alpha(1 - \mathrm{TrainAcc}) + \beta(1 - \mathrm{TestAcc}), \end{cases} \tag{4}$$

其中,N 为选中的特征个数,TrainAcc、TestAcc 分别为训练集和测试集正确率,在适应度函数中兼顾训练集和测试集的正确率,主要是防止训练过程中出现过拟合或者欠拟合。权重 α、β 取值为 0.4 和 0.6。

2.3 算法步骤

文中采用 5 折交叉验证评估算法的性能,算法最终的输出为最优参数集合和最佳特征组合子集。算法流程如图 1 所示:

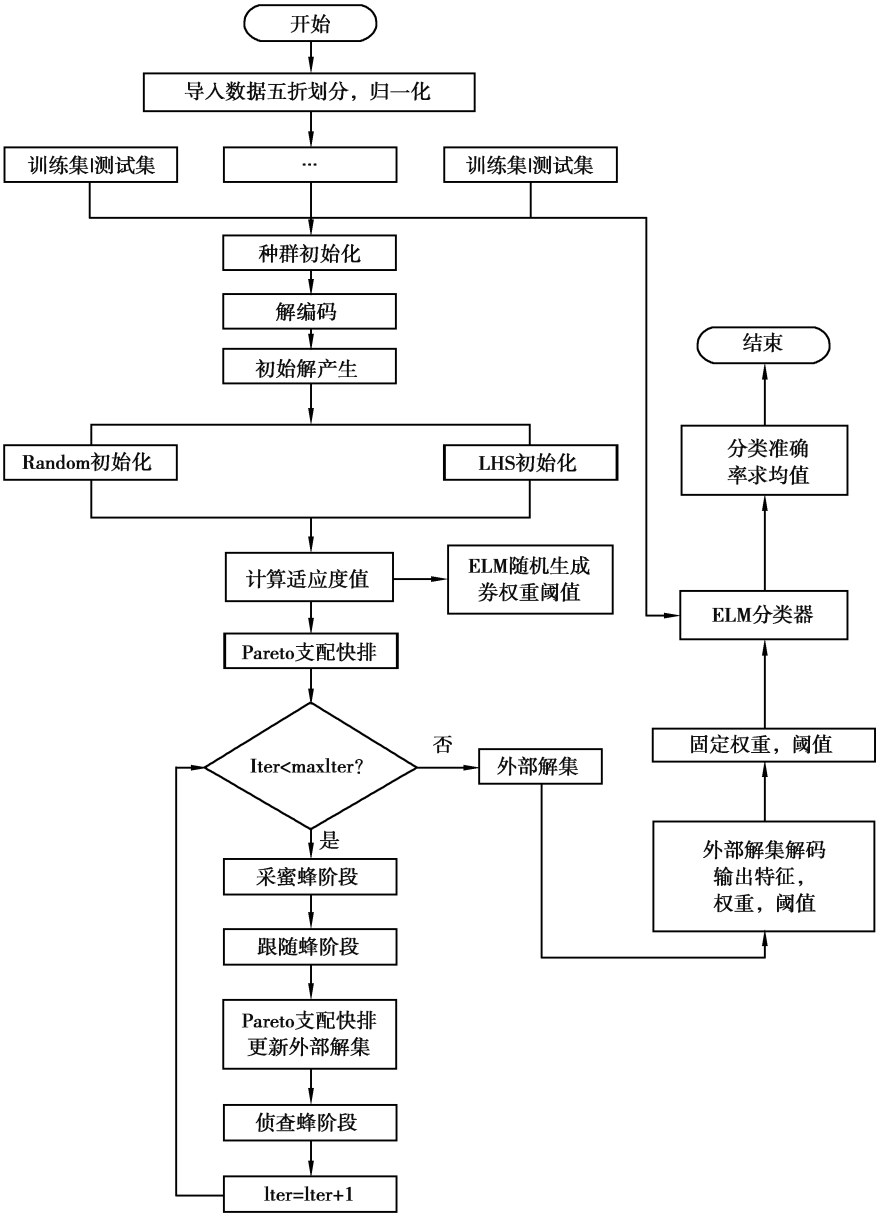

图 1 MOABC-ELM 算法流程图

Fig. 1 Flowchart of MOABC-ELM

步骤 1：归一化数据集，并按 5 折划分，其中训练集和测试集样本数比例为 4∶1。

步骤 2：初始化 MOABC 算法参数，生成初始解。针对 ABC 算法随机初始化方式(Random)产生的解分布不均匀问题，采用拉丁超立方体采样(LHS)方式生成初始解[12]。

步骤 3：计算初始解的适应度函数值，快速 Pareto 支配排序形成初始解集。

步骤 4：采蜜蜂阶段，更新解的维度得到新解，根据新解的适应度，确定新、旧解的 Pareto 支配关系，若旧解被新解支配，则存储新解，反之，旧解维持不变。

步骤 5：跟随蜂阶段，计算跟随蜂跟随采蜜蜂的概率。若随机生成的概率小于当前采蜜蜂的概率，则占用一个跟随蜂去采蜜蜂所在的解进行更新，过程和步骤 4 采蜜蜂阶段类似。

步骤 6：侦察蜂阶段：某一个解的领域搜索次数大于预设次数的蜜蜂角色转换为侦察蜂，并将所对应的解用随机生成的解替换。

步骤 7：当前搜索的最优解用外部档案进行存储。若外部档案中解的数目大于预设档案集数目，则根据拥挤距离来删除最劣解或者最拥挤的解。之后返回第 3 步继续循环，直到满足设定的搜索次数。

步骤 8：输出帕累托外部档案集，对其求平均值，并解码平均解生成的参数，将其带入模型，对优化后的分类器进行实验评估。

3　实验验证

3.1　实验数据描述

为了检验提出算法的分类性能，选用 UCI 机器学习数据库中的 6 个代表性数据集[13]进行测试，分别为 HAR、Wine、Iris、Breast Tissue、Statlog heart、Banknote，具体信息如表 1 所示。

表 1　实验数据集

Table 1　The experimental datasets

数据集	样本数	特征数	类别数
HAR	75 128	8	4
wine	178	13	3
Iris	150	4	3
Breast Tissue	106	9	6
Statlog heart	270	13	2
Banknote	1 372	4	2

3.2　实验结果及分析

算法运行在 Matlab 2016b 版本上，运行环境为 Windows7 操作系统、Intel core i5 处理器、主频 3.3 GHz、内存 4 GB。算法参数设置为种群规模 200、最大循环次数为 100、同一位置维度更新极限次数为 100，ELM 隐层激活函数为 Radbas。

为了确定隐含层节点数对分类结果的影响，这里设置节点数分别为 5、10、15、20、25、30，并分别考察不同的隐层神经元数目下 ELM 分类器的优化结果。由图 2 可以看出，随着节点数和特征个数的增加，错误率总体呈减小趋势，但并不一定是特征个数和节点数越多，错误率就越低。从结果出发，按照特征数和错误率两者均最小的原则，为不同的数据集选择各自 ELM 分类器的节点数目，结果如表 2 所示。

表 2 不同数据集的节点数设置

Table 2 The selection of nodes of ELM under different datasets

数据集	HAR	Wine	Iris	Banknote	Breast Tissue	Stagtlog heart
节点数	20	5	10	15	25	25

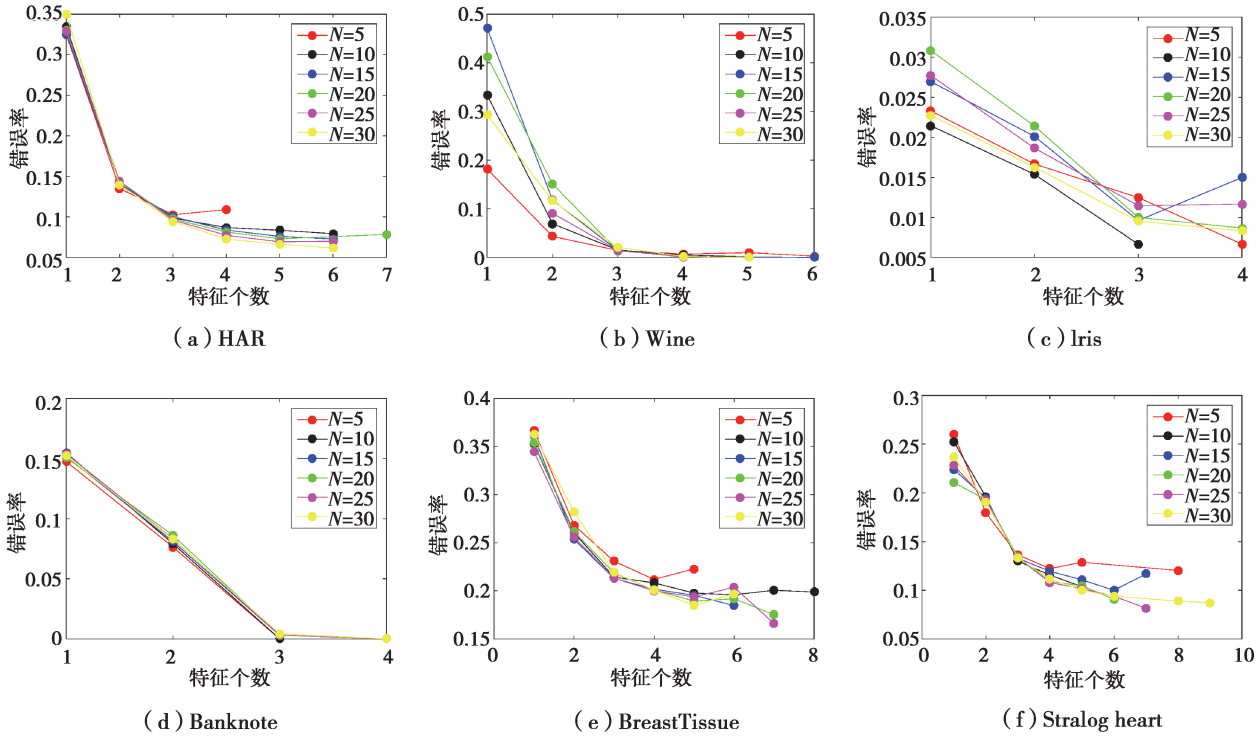

图 2 不同节点数下 Pareto 优化结果

Fig. 2 Theoptimization results with different nodes of ELM

为了评估算法的性能，基于选定的 ELM 神经元节点数，利用基于 LHS 初始化和传统随机初始化方法的人工蜂群算法去寻优特征个数和分类器参数，对生成的 Pareto 最优解所对应的适应度求平均值，其结果如图 3 所示。由图 3 可知，2 种初始化方法对应的多目标优化算法均能取得较好的效果。优化后的特征子集有多个选择，特征数目和错误率之间的权衡可以由用户自行选择。而当特征数目最少或最多时，分类错误率都不是最低。这也从一个侧面说明，解决这种问题使用多目标优化算法是最佳选择。

为了验证优化连接权值和阈值对结果产生的影响，图 4 列出了 2 种优化方式的 Pareto 结果图，其中蓝色折线(F_Pareto)代表只优化特征个数而手动选择 ELM 分类器参数的情况，而红色折线(FWB_Pareto)表示文中所采用的同时优化特征和 ELM 模型输入层和隐含层直间的连接权值、隐含层的阈值的方法。从图 4 可以看出，FWB_Pareto 方法相比 F_Pareto 方法，在特征个数相同的情况下，错误率更低，分类精度更高。证明了文中所提出方法的有效性。

为了更进一步验证文中方法的先进性，不考虑 Wine、Iris、Banknote 等 3 个数据集分类准确率为 100% 的情况，只在表 3 中列出另外 3 个数据集的分类结果，并与其它文献对应结果进行了对比。从表 3 可以看出，文中所用方法相比于参考文献的方法，能够在保证更少特征个数的前提下，有效地将正确率提升 5% 甚至更多。

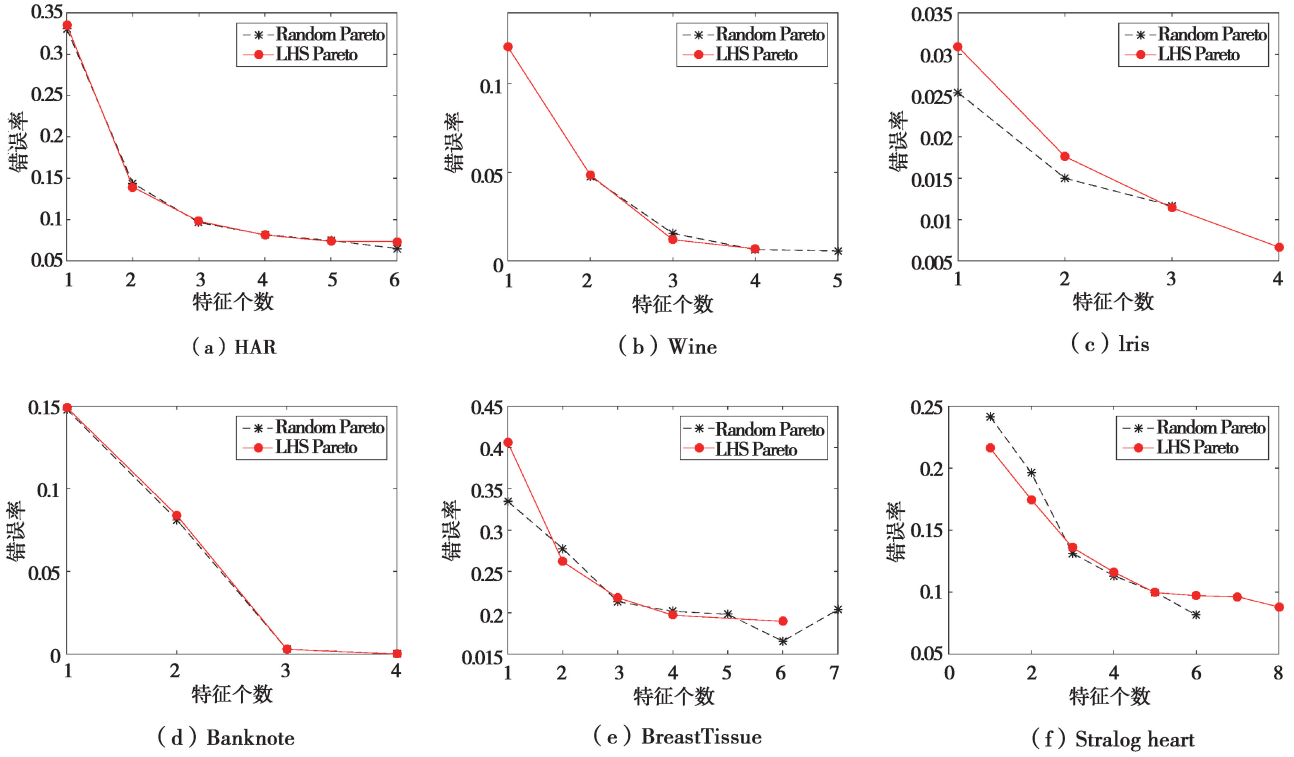

图 3　两种初始化方式结果对比

Fig. 3　Thecomparisons of results corresponding todifferent initialization methods

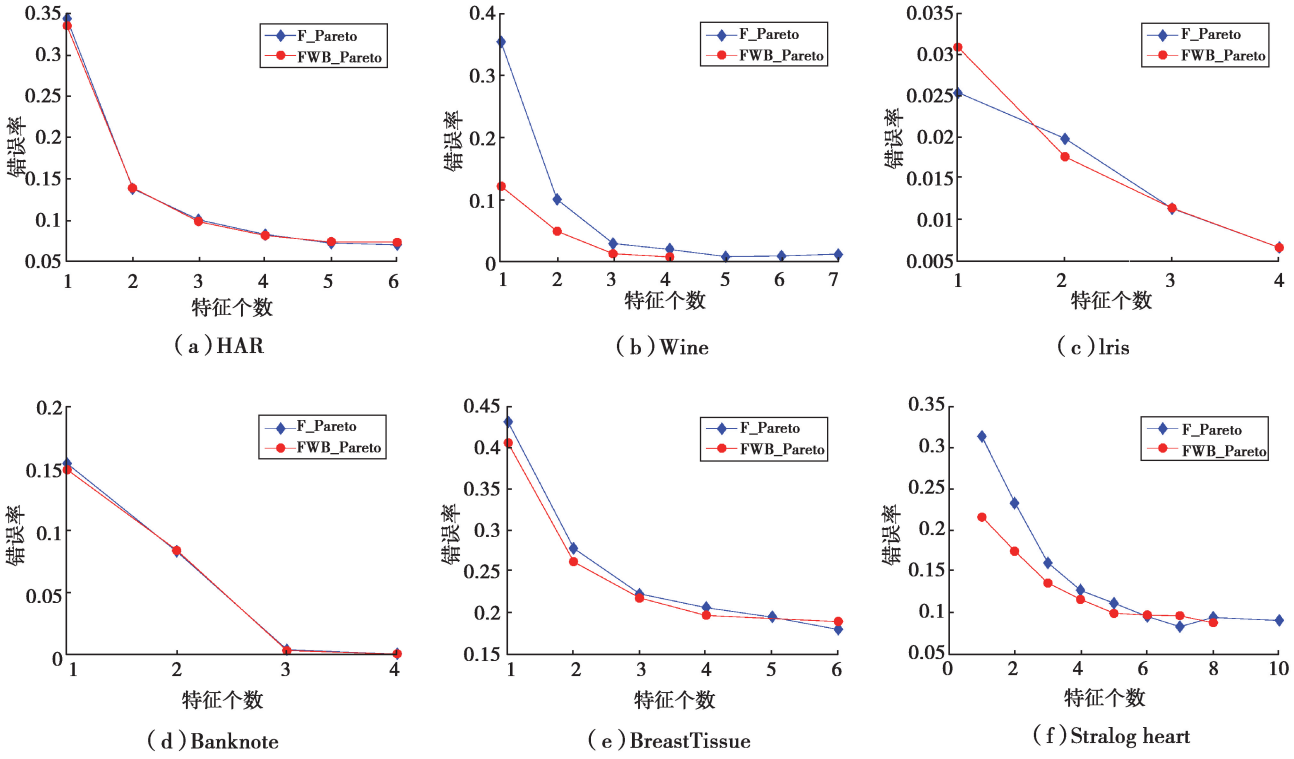

图 4　2 种优化方式下 Pareto 优化结果

Fig. 4　The optimization results with two methods

表3 文中与文献的结果对比

Table 3 The comparison of results in this paper and previous literatures

数据集	文献中正确率和特征数	文中正确率和特征数
HAR	85.5%,8[15]	90%,3
Breast Tissue	72%,9[16]	80%,3
Statlog heart	85%,7[17]	91%,7

4 结束语

文中将多目标蜂群算法应用于多维数据分类问题中,对如何同时完成特征选择和极限学习机的参数优化进行了探讨。通过UCI数据集的测试,特征个数少且分类错误率低,结果证明了所提出的方法的有效性。

参考文献:

[1] Han F, Zhao M R, Zhang J M. An improved incremental error minimized extreme learning machine for regression problem based on particle swarm optimization[J]. International Conference on Intelligent Computing,2015(8): 94-100.

[2] Han F, Zhao M R, Zhang J M. An improved incremental error minimized extreme learning machine for regression problem based on particle swarm optimization[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 94-100.

[3] 何敏,刘建伟,胡久松,遗传优化核极限学习机的数据分类算法[J].传感器与微系统,2017,36(10):141-143.
HE Min, LIU Jianwei, HU Jiusong. Genetic optimization kernel-based extreme learning machine data classification algorithm [J]. Transducer and Microsystem Technologies, 2017,36(10):141-143. (in Chinese)

[4] 赵虎, 左开伟, 覃永震. 改进人工蜂群算法优化ELM分类模型[J]. 计算机测量与控制, 2016, 24(10): 251-254.
ZHAO Hu, ZUO Kaiwei, QIN Yongzhen. Improved artificial bee colony optimize ELM classification model[J]. Computer Measurement & Control, 2016, 24(10): 251-254.(in Chinese)

[5] Bui D T, Ngo P T Thi, Pham T D, et al. A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping[J]. Catena, 2019, 179: 184-196.

[6] Krishnan G S, Sowmya K S. A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data [J]. Applied Soft Computing, 2019, 80: 525-533.

[7] Nayak D R, Dash R, Majhi B. Discrete ripplet-II transform and modified PSO based improved evolutionary extreme learning machine for pathological brain detection[J]. Neurocomputing, 2018, 282: 232-247.

[8] Wang M J, Chen H L, Li H Z, et al. Grey wolf optimization evolving kernel extreme learning machine: Application to bankruptcy prediction[J]. Engineering Applications of Artificial Intelligence, 2017, 63: 54-68.

[9] 盛晓晨,史旭东,熊伟丽. 改进粒子群优化的极限学习机软测量建模方法[J].计算机应用研究,2020,37(6).
SHENG Xiaochen, SHI Xudong, XIONG Weili. Soft sensor modeling for extreme learning machine based on improved particle swarm optimization[J]. Application Research of Computers, 2020,37(6).(in Chinese)

[10] 牛培峰, 李进柏, 刘楠,等. 基于改进花授粉算法和极限学习机的锅炉NO_x排放优化[J]. 动力工程学报, 2018, 38(10): 782-787.
NIU Peifeng, LI Jinbai, LIU Nan, et al. NO_x emission optimization of a boiler based on improved flower pollination algorithm and extreme learning machine[J]. Journal of Chinese Society of Power Engineering, 2018, 38(10): 782-787.(in Chinese)

[11] Huang G B, Zhou H, Ding X, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, 42(2): 513-529.

[12] Cui L Z，Li G H，Zhu Z X，et al. A novel artificial bee colony algorithm with an adaptive population size for numerical function optimization[J]. Information Sciences，2017，414：53-67.

[13] Evangelaras H，Koutras M V. On second order orthogonal Latin hypercube designs[J]. Journal of Complexity，2017，39：111-121.

[14] A Frank，Asuncion. UCI machine learningrepository.https：//archive.ics.uci.edu/ml/index.php

[15] Ling H，Qian C X，Kang W C，et al. Combination of Support Vector Machine and K-Fold cross validation to predict compressive strength of concrete in marine environment[J]. Construction and Building Materials，2019，206：355-363.

[16] Shinmoto Torres R L，Ranasinghe D C，Shi Q F. Evaluation of wearable sensor tag data segmentation approaches for real time activity classification in elderly[M] // Lecture Notes of the Institute for Computer Sciences，Social Informatics and Telecommunications Engineering. Cham：Springer International Publishing，2014：384-395.

[17] Liyao Ma，SebastienDestercke，Yong Wang. Online active learning of decision trees with evidential data[J]. Pattern Recognition，2016，52：33-45.

[18] Karegowda A G. Enhancing BPN Performance using GA identified significant features：a case study for categorization of heart statlog dataset[C].Foundation of Computer Science (FCS)，2013，IC2IT(1)：1-4.

（编辑　陈移峰）