doi:10.11887/j.cn.201806016 http://journal.nudt.edu.cn

稳定特征选择的多目标蚁群优化*

刘 艺[1],曹建军[2],刁兴春[1],张 磊[1]
(1. 陆军工程大学 指挥信息系统学院,江苏 南京 210007;2. 国防科技大学 第六十三研究所,江苏 南京 210007)

摘 要:为了提高进化算法特征选择稳定性,提出一种面向稳定特征选择的多目标蚁群优化方法。通过抽样策略集成三种特征排序法的输出作为多目标蚁群优化的稳定性指导信息,聚合特征的费舍尔分值和最大信息系数值作为多目标蚁群优化的启发式信息,以分类正确率和扩展昆彻瓦指标值作为两个优化目标,兼顾算法的分类性能与特征选择稳定性。在四个标准数据集上进行对比实验,结果表明,所提方法能够在分类性能与稳定性方面达到较好的平衡。

关键词:多目标蚁群优化;特征选择;特征选择稳定性;高维数据
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2018)06-118-06

Multiobjective ant colony optimization for stable feature selection

LIU Yi[1], *CAO Jianjun*[2], *DIAO Xingchun*[1], *ZHANG Lei*[1]
(1. College of Command Information Systems, PLA Army Engineering University, Nanjing 210007, China;
2. The 63rd Institute, National University of Defense Technology, Nanjing 210007, China)

Abstract: To improve the feature selection stability of evolutionary algorithms, a new method for stable feature selection based on multiobjective ant colony optimization was developed. Feature selection results of three feature ranking methods by resampling policy were combined to provide stable features' information for multiobjective ant colony optimization; the feature's Fisher discriminant value and maximal information coefficient value were integrated as heuristic information; the classification correctness rate and value of extensions of Kuncheva similarity measure were taken as two optimization objectives to balance algorithm's classification performance and its stability. Some comparisons and experiments were carried out on four benchmark data sets, and results show that the proposed method has a better tradeoff between classification performance and feature selection stability.

Key words: multiobjective ant colony optimization; feature selection; stability of feature selection; high dimensional data

随着大数据的发展和机器学习的广泛应用,在基因筛选、生物识别、癌症检测等领域,特征选择稳定性受到了广泛的关注[1]。特征选择稳定性是指特征选择方法对训练数据的微小扰动具有一定的鲁棒性,即能够提供较为稳定的特征集合。提升特征选择的稳定性,有助于发现相关特征,增强领域专家对结果的可信度,同时降低数据获取的时间耗费和复杂度[2]。

近年来,提升特征选择稳定性的研究已经成为热点,按所用技术是否与特征本身相关,特征选择稳定性提升方法可以分为扰动法和特征法。扰动法是在数据和特征选择算法层面的方法,它通过数据集进行扰动、增加新的数据实例以及将多种不同的特征选择方法进行集成等方式提升特征选择稳定性[3-5]。特征法是在特征层面进行进一步处理,然后与特定的特征选择方法相融合,提高特征选择的稳定性[6-7]。

传统研究只关注排序法的稳定性提升,对基于进化算法的特征选择稳定性提升研究则鲜见报道,提高进化算法特征选择稳定性已经成为亟待解决的关键问题[8]。

1 基于多目标蚁群优化的稳定特征选择

本文提出基于多目标蚁群优化的稳定特征选择(Stable Feature Selection based on MultiObjective Ant Colony Optimization, SFSMOACO)方法,该算法由3个主要部分构成,即集成特征排序、启发式信息生成和多目标蚁群优化(MultiObjective Ant Colony Optimization, MOACO),算法框架如图1所示。

* **收稿日期**:2017-09-15
基金项目:国家自然科学基金资助项目(61371196)
作者简介:刘艺(1990—),男,安徽蚌埠人,博士研究生,E-mail:albertliu20th@163.com;
曹建军(通信作者),男,副研究员,博士,硕士生导师,E-mail:jianjuncao@yeah.net

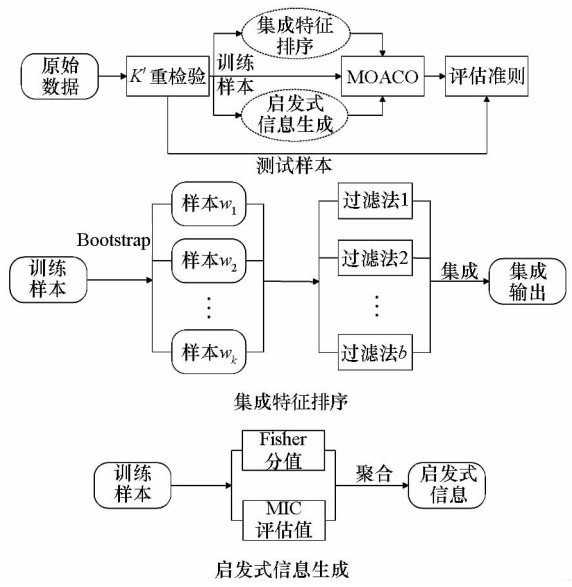


图 1　SFSMOACO 算法框架

Fig. 1　Framework of SFSMOACO

下面分别对算法的组成部分做详细的解释,并给出算法伪代码和时间复杂度分析。

1.1　MOACO

特征选择是典型的组合优化问题,MOACO 在解决组合优化问题方面具有较强的优越性。文献[9]提出了一种基于图的蚂蚁系统,使用等效路径信息素增强策略,具有较好的性能。但是该系统仅用于解决单目标组合优化问题,本文将其扩展到多目标优化问题中,作为 SFSMOACO 算法的重要组成,搜索较好特征子集。

在 MOACO 中,蚂蚁根据条件转移概率选择路径,其计算方法如式(1)所示。

$$p_{ij}^{k}=\begin{cases}\dfrac{[\tau_{ij}]^{\alpha}\cdot[\eta_i]^{\beta}}{\sum_{e_{uj}\notin visit_k}[\tau_{uj}]^{\alpha}\cdot[\eta_u]^{\beta}} & e_{ij}\notin visit_k\\ 0 & \text{otherwise}\end{cases}\tag{1}$$

式中:p_{ij}^{k}表示蚂蚁 k 从特征 d_j 经过边 e_{ij}到达特征 d_{j+1}的概率;τ_{ij}是边 e_{ij}在当前迭代中的信息素值;η_i 是静态启发式信息,表示特征 i 的选择期望;$visit_k$ 表示蚂蚁 k 访问过的边集;α 表示信息素值的重要程度;β 表示启发式信息的重要程度。

由于 SFSMOACO 同时优化分类性能和稳定性 2 个目标,因此在 MOACO 中对应优化目标设置 2 个信息素矩阵。为了计算条件转移概率,需要聚合信息素值,如式(2)所示。

$$\tau_{ij}=(\tau_{ij}^{1})^{(1-\lambda)}\cdot(\tau_{ij}^{2})^{\lambda}\tag{2}$$

式中,λ 是权重参数并且有 $0\leqslant\lambda\leqslant1$。

多目标优化问题的解称为帕累托解,帕累托解无优劣之分,SFSMOACO 设置档案保存帕累托解。使用档案中帕累托解更新 2 个信息素矩阵,更新方法见式(3)。

$$\tau_{ij}(t)=\begin{cases}H(t)+\Delta'(tabu^{t}) & e_{ij}\in\Gamma(tabu^{t})\\ H(t) & \text{otherwise}\end{cases}\tag{3}$$
$$H(t)=(1-\rho)\tau_{ij}(t-1)$$

其中,ρ 是信息素挥发系数,$\Delta'(tabu^{t})$是信息素更新增量,$\Gamma(tabu^{t})$是等效路径。MOACO 将当前更新候选解在目标上的适应度值作为信息素更新增量,加强较好路径对算法的影响,提高算法获取较好解的能力。$\Delta'(tabu^{t})$通过式(4)计算。

$$\Delta'(tabu^{t})=[\sum_{h=1}^{m}f_h(tabu^{t})]/(Q\times m)\tag{4}$$

式中,m 表示目标的个数,$f_h(tabu^{t})$表示路径 $tabu^{t}$ 在目标空间中第 h 个目标上的值,Q 是常数。蚁群算法其他相关部分见文献[9]。

1.2　集成特征排序

研究表明,通过集成多种特征排序法的结果能够有效提升特征选择稳定性[10]。文献[11]通过实验证明,信息增益和卡方检验具有较好的稳定性,ReliefF 同时具有较好的分类性能和稳定性,不同的集成策略并无显著的差异。因此本文选择这 3 种特征选择方法对抽样数据进行排序,并采用中值法集成多组特征排序列表,即将某个特征在多个特征排序列表中的中间排序结果作为该特征的最终排序。

1.3　启发式信息生成

在 MOACO 算法中,启发式信息表示蚂蚁在路径选择中的先验偏好,它与具体的问题相关,定义合适的启发式信息能够有效提升算法获得较好解的能力。本文基于特征选择的应用背景,提出采用 Fisher 分值与最大信息系数(Maximal Information Coefficient, MIC)评估值相结合的启发式信息定义方法。利用 Fisher 分值选择具有较强判别能力的特征,使用 MIC 得出稳定的特征排序,通过集成 Fisher 分值与 MIC 评估值,提高 SFSMOACO 选择同时具有较好分类性能与稳定性的特征子集的概率。

Fisher 分值是一种度量特征判别能力的指标,以二分类问题为例,在训练样本中,第 h 个特征的 Fisher 分值采用式(5)计算。

$$Fscore(h)=\frac{|\bar{\mu}_{1h}-\bar{\mu}_{2h}|}{\sqrt{\sigma_{1h}^{2}+\sigma_{2h}^{2}}}\tag{5}$$

式中:$\bar{\mu}_{1h}$,$\bar{\mu}_{2h}$分别为正类和负类内第 h 个特征值的均值;σ_{1h}^2,σ_{2h}^2分别为正类和负类内第 h 个特征的方差。Fisher 分值越大,特征的判别能力越强。

MIC 的内涵是,如果两个变量之间存在某种关系,那么可以通过某种方法在由它们构成的空间中给出网格,从而使得一些单元格包含大部分的点[12]。

假设数据集中的变量构成集合 D,每个实例分布在 D 的散点图上,将实例按 x 值划分到 x 个单元格,同时将实例按 y 值划分到 y 个单元格,这种方式称为 $x \times y$ 划分。D 中的点在网格 G 上分布的概率为$D|_G$,则当确定集合 D 时,G 将确定唯一的概率分布。

定义 1 对集合 $D \subset \mathbb{R}^2$ 和两个正整数 x、y,有:

$$I^*(D,x,y) = \max I(D|_G) \tag{6}$$

式中,$I(D|_G)$表示集合中的点在网格 G 中的分布 $D|_G$ 的互信息,最大值为在所有 G 中取值的最大值。

定义 2 D 中的特征矩阵元素是:

$$M(D)_{x,y} = \frac{I^*(D,x,y)}{\log_2 \min\{x,y\}} \tag{7}$$

式中,$M(D)_{x,y}$是一个无穷阵。

定义 3 假设 D 有 n 个实例,同时网格数的最大值为 $B(n)$,则其 MIC 值是:

$$MIC(D) = \max_{xy<B(n)}\{M(D)_{x,y}\} \tag{8}$$

MIC 的取值是(0, 1),MIC 的评估值越大,表明变量之间的相关性越强,同时有 $MIC(X, Y) = MIC(Y, X)$。

评估特征相关性时,将该特征取值与对应的类标用两个向量表示,采用 MIC 计算两个向量的值,得出该特征与类别的相关性。

1.4 优化目标

SFSMOACO 使用分类正确率作为分类性能指标,采用扩展昆彻瓦指标度量特征子集的稳定性[13]。

设参与分类的样本总数为 N_m,被正确区分为正类的样本数为 P_n,被正确区分为负类的样本数为 N_n,则分类正确率 P 计算如式(9)。

$$P = \frac{P_n + N_n}{N_m} \tag{9}$$

设由同一特征选择方法生成的两个特征子集为 s 和 s',它们的扩展昆彻瓦指标值 EK 的计算如式(10)所示。

$$\begin{cases} EK(s,s') = \dfrac{|s \cap s'| - \Upsilon}{\max[\Phi^1;\Phi^2]} \\ \Phi^1 = -\max(0,|s| + |s'| - c) + \Upsilon \\ \Phi^2 = \min(|s|,|s'|) - \Upsilon \\ \Upsilon = \dfrac{|s| \cdot |s'|}{c} \end{cases} \tag{10}$$

其中,c 表示数据集中特征的个数。指标 EK 的取值范围是[-1, 1],取值越大两个特征子集越相似,算法的稳定性越好。采用扩展昆彻瓦指标度量 MOACO 选择的特征子集与集成特征排序结果之间的相似度,以集成特征排序信息指导 MOACO 的进化过程,使得算法能够选择稳定的特征子集。

1.5 算法描述

综上所述,本文提出的 SFSMOACO 算法的伪代码如算法 1 所示。

算法 1 SFSMOACO 伪代码

Alg. 1 Pseudo-code of SFSMOACO

输入:原始数据,算法相关参数

输出:特征子集和数据分类结果

```
1. begin
2. for K′重交叉检验
3.     对训练样本进行 Bootstrap 抽样生成 k 组抽样数据
4.     在 k 组抽样数据上采用 3 种特征选择方法进行特征排序,并融合排序结果
5.     计算训练样本中每个特征的 Fisher 分值和 MIC 评估值作为启发式信息
6.     以 P 和 EK 为优化目标,采用 MOACO 搜索较好特征子集
7. end for
8. end
```

现对 SFSMOACO 算法的时间复杂性做分析。设数据维度为 C,信息增益、卡方检验和 ReliefF 的时间复杂度为 $O(k \times C)$,计算 Fisher 分值与 MIC 评估值的时间复杂度为 $O(C)$;设 MOACO 的蚂蚁个数为 N_a,最大迭代次数为 N_c,则 MOACO 的时间复杂度为 $O(N_c \times N_a \times C^2)$。综上,可以得出 SFSMOACO 算法的时间复杂度为 $O(N_c \times N_a \times C^2 + k \times C + C)$。

2 实验与分析

实验使用典型的二分类高维数据集,数据集来源于亚利桑那州立大学特征选择算法与数据档案网站[14],实验数据的属性如表 1 所示。

表 1　实验数据集属性

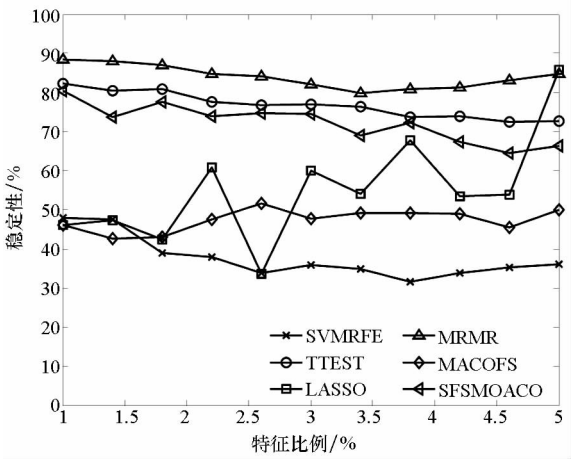
Tab. 1　Characteristics of experiment datasets

数据集	简称	实例数	特征数
PCMAC	PC	1943	3289
RELATHE	RE	1427	4322
BASEHOCK	BA	1993	4862
MADELON	MA	2600	500

采用 T 检验方法 TTEST[3]、支持向量机递归特征消除[15](Support Vector Machine Recursive Feature Elimination,SVMRFE)、最小绝对值收敛和选择算子[16](Least Absolute Shrinkage and Selection Operator, LASSO)、最小冗余最大相关[17](Minimal Redundancy Maximal Relevance, MRMR)和文献[18]提出的基于多目标蚁群优化的特征选择(Multiobjective Ant Colony Optimization based Feature Selection, MACOFS)5 种方法作为对比算法。TTEST、SVMRFE、LASSO、MRMR 在分类性能与稳定性方面具有较好的综合性能;与 MACOFS 相比,SFSMOACO 的主要区别在于引入了 MIC 特征评估作为启发式信息,因此有必要通过实验验证 SFSMOACO 算法在稳定性与分类性能上的优越性。

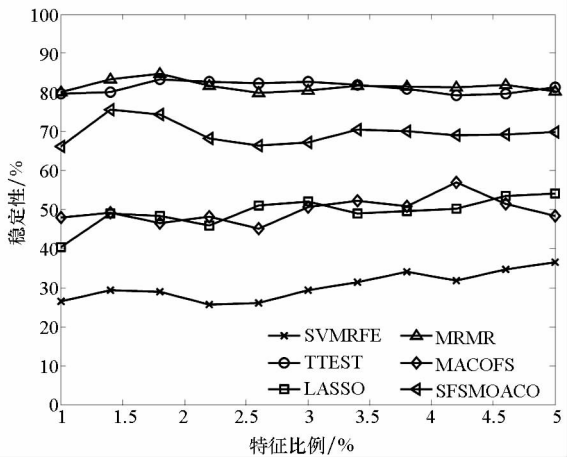
SFSMOACO 的参数设置为:$k=11$,$\alpha=1$,$\beta=2$,$Q=0.1$,信息素初始浓度 $\tau(0)=100$,$\lambda=0.5$,迭代次数为 200,帕累托档案规模为 40。实验采用支持向量机作为分类器,并将 20 轮 5 重交叉检验(即 $K'=5$)的均值作为输出。特征子集规模占全部特征比例变化设置从 1% ~5%,SFSMOACO 在 4 个数据集上的扩展昆彻瓦指标值的对比结果如图 2 所示。

从图 2 可以看出,在 PC、RE 和 BA 数据集



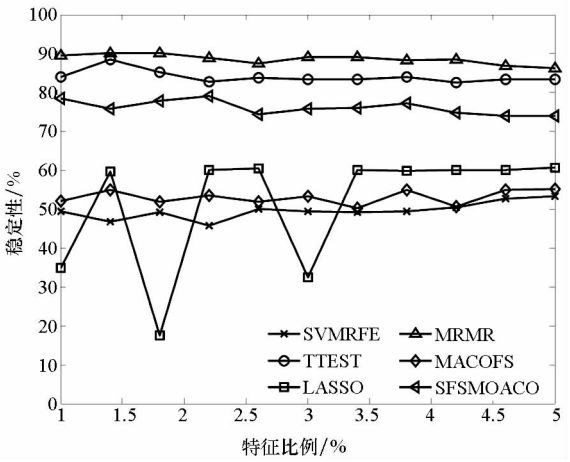
(a) SFSMOACO 在 PC 中的结果

(a) Results of SFSMOACO on PC



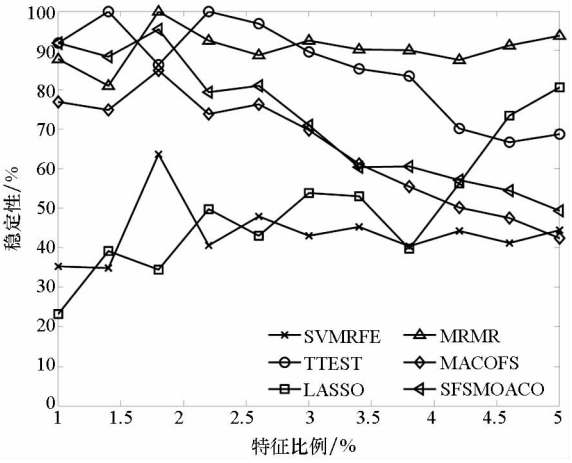
(b) SFSMOACO 在 RE 中的结果

(b) Results of SFSMOACO on RE



(c) SFSMOACO 在 BA 中的结果

(c) Results of SFSMOACO on BA



(d) SFSMOACO 在 MA 中的结果

(d) Results of SFSMOACO on MA

图 2　SFSMOACO 的特征选择稳定性结果

Fig. 2　Stability results of SFSMOACO

上,SFSMOACO 方法的稳定性要弱于 MRMR 和 TTEST 方法的,优于 LASSO、SVMRFE、MACOFS 方法的,并且相比 SVMRFE、LASSO 和 MACOFS 方法,SFSMOACO 方法的稳定性变化趋势较为稳

定,说明 SFSMOACO 在特征选择稳定性方面具有较好的鲁棒性。在 MA 数据集上,特征比例小于4.5%时,SFSMOACO 方法的稳定性优于 SVMRFE 和 LASSO 2 种方法的,仅在特征比例大于 4.5%时,LASSO 方法的稳定性优于 SFSMOACO。从 4 个子图中可以看出,SFSMOACO 的稳定性在所有测试数据集上都要优于 MACOFS 方法。从图2(d)中可以看出,SFSMOACO 方法的稳定性随着特征比例的增加呈现下降趋势,而 MACOFS 亦有相同变化趋势,可以推断,造成 SFSMOACO 稳定性下降的原因可能是 MOACO 的性能恶化。

SFSMOACO 的分类正确率比较实验结果如表2~5所示。

表2 SFSMOACO 在 PC 中的分类正确率

Tab.2 Classification success rate of SFSMOACO on PC

算法	特征比例				
	1%	2%	3%	4%	5%
SFSMOACO	**0.96**	**0.95**	**0.94**	**0.92**	**0.92**
SVMRFE	0.89	0.89	0.88	0.90	0.89
TTEST	0.87	0.89	0.90	0.89	0.90
LASSO	0.79	0.81	0.81	0.83	0.81
MRMR	0.87	0.88	0.89	0.90	0.91
MACOFS	0.90	0.91	0.92	0.91	0.92

表3 SFSMOACO 在 RE 中的分类正确率

Tab.3 Classification success rate of SFSMOACO on RE

算法	特征比例				
	1%	2%	3%	4%	5%
SFSMOACO	**0.94**	**0.94**	**0.92**	0.92	0.92
SVMRFE	0.84	0.84	0.85	0.86	0.85
TTEST	0.76	0.82	0.85	0.86	0.86
LASSO	0.74	0.81	0.82	0.81	0.83
MRMR	0.76	0.81	0.84	0.86	0.85
MACOFS	0.89	0.91	**0.92**	**0.94**	**0.94**

表4 SFSMOACO 在 BA 中的分类正确率

Tab.4 Classification success rate of SFSMOACO on BA

算法	特征比例				
	1%	2%	3%	4%	5%
SFSMOACO	**0.95**	**0.95**	0.94	0.94	0.93
SVMRFE	0.94	**0.95**	0.96	0.96	0.96
TTEST	0.93	**0.95**	0.96	**0.97**	0.96
LASSO	0.59	0.67	0.78	0.80	0.79
MRMR	0.94	**0.95**	0.96	0.87	0.86
MACOFS	**0.95**	**0.95**	**0.97**	**0.97**	**0.98**

表5 SFSMOACO 在 MA 中的分类正确率

Tab.5 Classification success rate of SFSMOACO on MA

算法	特征比例				
	1%	2%	3%	4%	5%
SFSMOACO	**0.98**	**0.99**	**0.93**	0.86	0.81
SVMRFE	0.50	0.51	0.55	0.56	0.55
TTEST	0.54	0.52	0.55	0.50	0.51
LASSO	0.50	0.50	0.51	0.52	0.48
MRMR	0.53	0.54	0.54	0.49	0.52
MACOFS	0.90	0.90	0.91	**0.91**	**0.91**

从对比方法的角度分析:SFSMOACO 和 MACOFS 在多数的测试条件下取得了较好的效果。在 PC 数据上,SFSMOACO 方法提供了全部5个最优解,在 RE 数据集上提供了3个最优值,BA 数据上取得2个最优解,在 MA 数据上获得了3个最优值。可以看出,与 MACOFS 方法相比,SFSMOACO 方法的分类性能更好。

从稳定性与分类性能平衡的角度分析:虽然 SFSMOACO 方法的稳定性弱于 TTEST 和 MRMR 2种方法的,但要优于 MACOFS 方法的。在分类性能上,SFSMOACO 方法的分类正确率在多数情况下要好于 TTEST、MRMR 和 MACOFS 3种方法的;而 LASSO 和 SVMRFE 2种方法在稳定性上弱于 SFSMOACO,在分类性能上也要弱于 SFSMOACO。因此可以得出结论,SFSMOACO 方法能够在稳定性与分类性能上达到较好的平衡,同时其综合性能也好于 MACOFS 的。

3 结论

提高特征选择稳定性是特征选择的重要研究内容,本文提出一种提高多目标蚁群进化算法特征选择稳定性的 SFSMOACO 方法,集成3种特征排序方法结果作为稳定性指导信息,结合 Fisher 分值与 MIC 特征评估值作为 MOACO 的启发式信息。经过实验对比分析得出结论,SFSMOACO 方法能够在特征选择稳定性与分类性能上达到较好的平衡,在稳定性方面具有较好的鲁棒性,同时也表明引入 MIC 特征评估作为 MOACO 启发式信息的有效性。

下一步,将继续探究特征选择稳定性与分类性能之间的关系及影响要素,同时分析影响进化算法本身稳定性的因素,进一步提升 SFSMOACO 方法的综合性能。

参考文献(References)

[1] Kim H J, Choi B S, Huh M Y. Booster in high dimensional data classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(1): 29 – 40.

[2] Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high dimensional spaces [J]. Knowledge and Information Systems, 2007, 12 (1): 95 – 116.

[3] Zhou Q F, Ding J C, Ning Y P, et al. Stable feature selection with ensembles of multi-ReliefF[C]//Proceedings of the 10th International Conference on Natural Computation, 2014: 742 – 747.

[4] Rondina J M, Hahn T, Oliveira L D, et al. SCoRS: a method based on stability for feature selection and mapping in neuroimaging[J]. IEEE Transactions on Medical Imaging, 2014, 33(1): 85 – 98.

[5] Fahad A, Tari Z, Khalil I, et al. An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion [J]. Future Generation Computer Systems, 2014, 36: 156 – 169.

[6] Shu L, Ma T Y, Latecki L J. Stable feature selection with minimal independent dominating sets [C]//Proceedings of ACM International Conference on Bioinformatics, 2013: 450 – 457.

[7] Beinrucker A, Dogan U, Blanchard G. Extensions of stability selection using subsamples of observations and covariates[J]. Statistics and Computing, 2016, 26(5): 1059 – 1077.

[8] Xue B, Zhang M J, Brownw W N, et al. A survey on evolutionary computation approaches to feature selection[J]. IEEE Transactions on Evolutionary Computation, 2016, 20(4): 606 – 626.

[9] 曹建军, 张培林, 王艳霞, 等. 一种求解子集问题的基于图的蚂蚁系统[J]. 系统仿真学报, 2008, 20(22): 6146 – 6150.
CAO Jianjun, ZHANG Peilin, WANG Yanxia, et al. A graph-based ant system for subset problems[J]. Journal of System Simulation, 2008, 20(22): 6146 – 6150. (in Chinese)

[10] Kamker I, Gupta S K, Phung D, et al. Stabilizing l 1-norm prediction models by supervised feature grouping[J]. Journal of Biomedical Informatics, 2016, 59: 149 – 168.

[11] Pes B, Dessi N, Angioni M. Exploiting the ensemble paradigm for stable feature selection: a case study on high dimensional genomic data [J]. Information Fusion, 2017, 35: 132 – 147.

[12] Zhang Y, Jia S L, Huang H Y, et al. A novel algorithm for the precise calculation of the maximal information coefficient [J]. Scientific Reports, 2014(4): 6662.

[13] Nogueria S, Brown G. Measuring the stability of feature selection with applications to ensemble methods [C]// Proceedings of Multiple Classifier Systems: 12th International Workshop, 2016: 135 – 146.

[14] Arizona State University. Feature selection datasets [EB/OL]. [2017 – 08 – 15]. http://featureselection. asu. edu/datasets. php.

[15] Guyon I, Weston J, Barnhil S, et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46(1/2/3): 389 – 422.

[16] Shi W L, Wahba G, Irizarry R, et al. The partitioned LASSO-Patternsearch algorithm with application to gene expression data[J]. BMC Bioinformatics, 2012, 13: 98.

[17] Peng H C, Long F H, Ding C. Feature selection based on mutual information criteria of max dependency, max relevance, and min redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226 – 1238.

[18] Liu Y, Diao X C, Cao J J, et al. Evolutionary algorithms' feature selection stability improvement system [C]// Proceedings of the 12th International Conference on Bio-inspired Computing: Theories and Applications, 2017.