计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(7):145−151 [doi: 10.15888/j.cnki.csa.007463]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

参数自动优化的特征选择融合算法①

吴　俊, 柯颺挺, 任　佳

(浙江理工大学 机械与自动控制学院, 杭州 310018)
通讯作者: 任　佳, E-mail: jren@zstu.edu.cn

摘　要: 针对传统特征选择方法如信息增益存在选择偏好、处理非线性问题能力弱、以及参数手动优化过程繁琐的问题, 提出一种基于最大互信息系数与皮尔逊相关系数的两阶段特征选择融合算法, 并利用遗传算法对其中两个超参数自动进行优化. 第一阶段, 利用最大互信息系数获取特征和标签之间的相关性来进行特征选择; 第二阶段, 使用皮尔逊相关系数对获取的特征子集进行去冗余. 进一步, 基于遗传算法对两个阶段中的两个超参数自动进行优化. 将该方法运用于多组 UCI 数据集中进行测试. 实验结果表明, 该算法能够兼顾降低特征空间的维度和提升算法的分类性能.

关键词: 最大互信息系数; 皮尔逊相关系数; 特征选择; 遗传算法; 参数优化

引用格式: 吴俊,柯颺挺,任佳.参数自动优化的特征选择融合算法.计算机系统应用,2020,29(7):145–151. http://www.c-s-a.org.cn/1003-3254/7463.html

Parameter Automatic Optimization for Feature Selection Fusion Algorithm

WU Jun, KE Liu-Ting, REN Jia

(Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In view of traditional feature selection methods such as information gain algorithm have preference for selecting features that have more values, Pearson correlation coefficient alone cannot be used to deal with nonlinear correlation, and optimization of algorithm parameters is too tedious, a feature selection fusion approach is proposed based on maximum information coefficient and Pearson correlation coefficient. Moreover, this approach makes use of genetic algorithm to optimize parameters automatically. In the first stage, the feature selection is carried out according to the maximum information coefficient and the correlation between features and tags. In the second stage, Pearson correlation coefficient is used to reduce the redundant acquired features. Furthermore, two hyper-parameters in the first two stages are optimized automatically based on genetic algorithm. The experimental results show that the algorithm can reduce the dimension of feature space and improve the classification performance.

Key words: maximum information coefficient; Pearson correlation coefficient; feature selection; genetic algorithm; parameter optimization

随着现代科技的飞速发展, 大量数据从各个领域中呈爆炸式不断产出. 同时因为这些数据中含有大量不相关、冗余信息, 在进行数据挖掘时, 预处理技术中的特征选择便变得极为重要和极具挑战性[1].

目前特征选择已经在文本分类[2]、模式识别[3−5]、癌症分类[6]和故障诊断[7]等领域内成功应用. 特征选择可定义为从数据集中去除不相关和冗余的特征, 从而增强后续学习算法性能的过程[8]. 特征选择方法可根据两个标准来分类: 搜索策略和评估标准[8]. 特征子集可以通过两种模型获取: 过滤式[9]模型和封装式[10]模型.

① 基金项目: 浙江省自然科学基金 (LY17F030024); 浙江省公益技术研究项目 (GG20F030031)
Foundation item: Natural Science Foundation of Zhejiang Province (LY17F030024); Technology Research Project for Public Welfare of Zhejiang Province (GG20F030031)
收稿时间: 2019-11-19; 修改时间: 2019-12-11; 采用时间: 2019-12-25; csa 在线出版时间: 2020-07-03

过滤式模型在搜索过程中仅考虑数据集本身, 而封装式模型需在搜索过程中结合学习算法. 因此过滤式模型花费时间成本较小, 但准确率不稳定. 反观封装式模型, 由于需要结合学习算法, 所以准确率相对更高, 同时也更耗时. 此外搜索最优子集的方式也是特征选择中的一个关键问题, 具体可分为遍历、随机搜索和启发式搜索. 其中遍历虽然一定可以获取最优子集但是不适用于大型数据集[11]. 随机搜索则因为没有数学理论引导搜索方向导致搜索性能不稳定[12]. 而启发式算法使用启发式信息指导搜索. 虽然不能保证找到最优解, 但可以再合理的时间内获取可接受的解[13]. 近年已经有许多研究对不同特征选择算法进行了对比, 如在文献[14]中, 使用了 3 种不同的分类器评估了 8 种标准的特征选择方法. 文献[10]使用了 K 最近邻算法作为分类器, 系统地研究了鲸鱼优化算法 (Whale Optimization Algorithm, WOA) 并将基于 WOA 的封装式特征选择算法和三种标准的启发式特征选择算法进行比较证明了 WOA 的强大性能. 文献[15]中将融合启发式算法和 3 种标准的封装式特征选择算法进行了对比以证明融合启发式算法的潜力.

综上, 本文基于过滤式模型和封装式模型各自的特点, 将这两种模型结合, 构建一种基于最大互信息系数和皮尔逊相关系数的、使用遗传算法进行超参数自动优化的两阶段特征选择融合算法 (A feature selection fusion method based on Maximal Information Coefficient and Pearson correlation coefficient with parameters automatic optimized by Genetic Algorithms, MICP-GA), 通过结合使用最大互信息系数、皮尔逊相关系数的过滤式模型和以分类准确率为目标的封装式模型来克服过滤式模型准确率较低, 传统相关系数处理非线性关系能力较弱以及封装式模型的时间成本过高的问题. 第一阶段根据最大互信息系数获取各特征和标签之间的相关度评分, 该评分综合考虑线性和非线性相关度, 再设置相关度阈值以剔除不相关特征. 第二阶段, 通过皮尔逊相关系数获取特征子集特征之间的线性冗余度评分, 同样设置冗余度阈值来删除冗余特征. 最后, 将特征子集的分类准确率作为评价标准, 使用遗传算法自动优化前两步中的超参数, 达到综合减少特征数目和维持甚至提高特征子集分类精度的效果, 并自动获取最优特征子集.

1 基础理论

1.1 信息熵与最大互信息系数

定义 1. 信息熵[16]克服了对信息随机变量不确定性的度量, 设 X 为离散随机变量, 那么 X 的信息熵 $H(X)$ 为:

$$H(X) = -\sum_{i=1}^{m} p(x_i)\log p(x_i) \tag{1}$$

定义 2. 条件熵表示为当随机变量 Y 单独发生时, 随机变量 X 发生的条件概率分布, 可通过式 (2) 表示:

$$H(X|Y) = \sum_{y\in Y} p(y)\sum_{x\in X} p(x|y)\log p(x|y) \tag{2}$$

定义 3. 互信息用于检测两随机变量中所含的信息量和互相关联程度. 互信息可通过式 (1) 和式 (2) 用熵表示:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{3}$$

最后依据式 (2) 和式 (3) 可得:

$$0 \le I(X;Y) \le \min\{H(X), H(Y)\} \tag{4}$$

Reshef 等人[17]提出的最大互信息系数 (Maximal Information Coefficient, MIC) 无需对数据分布进行任何假设便可评估变量间的函数关系和统计关系. 该算法具有普适性和均匀性两大特点, 普适性指当数据规模足够大时, MIC 算法能有效捕捉到大规模有意义的关系, 而并不会仅局限于某种函数关系; 均匀性则指对于不同类型函数关系, 当给予相同噪声时, MIC 算法给出相同或相近的结果变化. 最大互信息系数 $I_{\max}(X;Y)$ 可通过互信息和熵计算得出:

$$I_{\max}(X;Y) = \frac{I(X;Y)}{\min\{H(X), H(Y)\}} \tag{5}$$

1.2 皮尔逊相关系数

皮尔逊相关系数 (Pearson Correlation Coefficient, PCCs) 由 Karl Person 提出, 因其计算简单、运算速度快而被广泛用于度量在数据预测、故障诊断和参数估计等领域中两变量间的线性相关程度, 其取值范围是[−1,1]. 两变量间相关系数的绝对值越大, 表明两者的相关度越高, 当取值为 0 时表示两个变量不相关. 具体相关系数数值大小和相关判断结果的对应关系见表 1.

通常皮尔逊相关系数用希腊字母 ρ 表示. 其具体定义为两个变量之间的协方差和标准差的商, 其中 cov 表示协方差, E 为数学期望:

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E((X-\mu_X)(Y-\mu_X))}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(X^2)-E^2(Y)}} \tag{6}$$

表 1 相关性判断准则

相关系数值 (绝对值)	判断结果
1.0	完全相关
0.7–1.0	强相关
0.4–0.7	弱相关
0.0–0.4	不相关

1.3 遗传算法

遗传算法的灵感来自生物的遗传过程, 该算法的解被称为染色体或个体. 二进制遗传算法的每个染色体包含二进制值为 0 或 1 的基因, 这些基因决定了每个个体的属性. 一系列的染色体组成了一个种群. 每个染色体的性能通过适应度函数来评估, 适应度值较高的染色体被选作父代, 并通过交叉步骤结合产生新的后代. 再对新的种群进行变异处理来增加个体的随机性, 降低陷入局部最优的可能[18].

2 MICP-GA 基础理论

MICP-GA 算法兼顾过滤式和封装式特征选择算法的优点, 同时利用遗传算法对前述步骤中的超参数进行自动优化来获取最优特征子集. 该算法的实现流程如图 1 所示.

2.1 运用最大互信息系数获取各特征和标签的相关度

本阶段利用最大互信息系数进行初次特征选择. 第一阶段特征选择的具体步骤如算法 1 所示.

该方法使用最大互信息系数准确找出和类别线性相关、非线性相关的特征, 剔除不相关的特征, 但筛选出的特征之间仍可能存在线性冗余. 为解决该问题, 继续进行第二步特征选择.

2.2 基于皮尔逊相关系数的特征去冗余

对第一步选取的特征子集 $D2$, 使用皮尔逊相关系数, 减少特征之间的冗余性, 同时也减少了 $D2$ 的特征数量, 帮助后续学习算法更快地获取结果. 第二阶段特征选择的步骤如算法 2 所示.

由前述可知, 第一阶段中的超参数 a 和第二阶段中的阈值 b 这两个参数值对分类结果有着重要的影响. 接下来, 以 UCI 标准数据集中的 Vehicle 数据集为例进行说明: 该数据集初始特征共计 18 维, 如在第一阶段特征选择时设定 a 为 18, 则保留全部特征. 现对所选的 Vehicle、Ionosphere、Wine 和 Sonar 数据集中的特征的皮尔逊相关系数进行可视化处理, 如图 2~图 5 所示. 热力图中, 两两特征之间的皮尔逊相关系数的取值范围是[0,1], 对应着图中不同的颜色. 此外, 由于热力图沿着正对角线 (如图 2~图 5 中从左上贯穿至右下的蓝线所示) 对称, 所以仅观察热力图的左下部分即可. 由图可知, 其中若干区域 (如图 2~图 5 中蓝色圈注所示) 中特征之间的皮尔逊相关系数较大, 说明该区域内的特征之间存在线性冗余. 可采用基于皮尔逊相关系数的特征选择算法有效解决该问题.

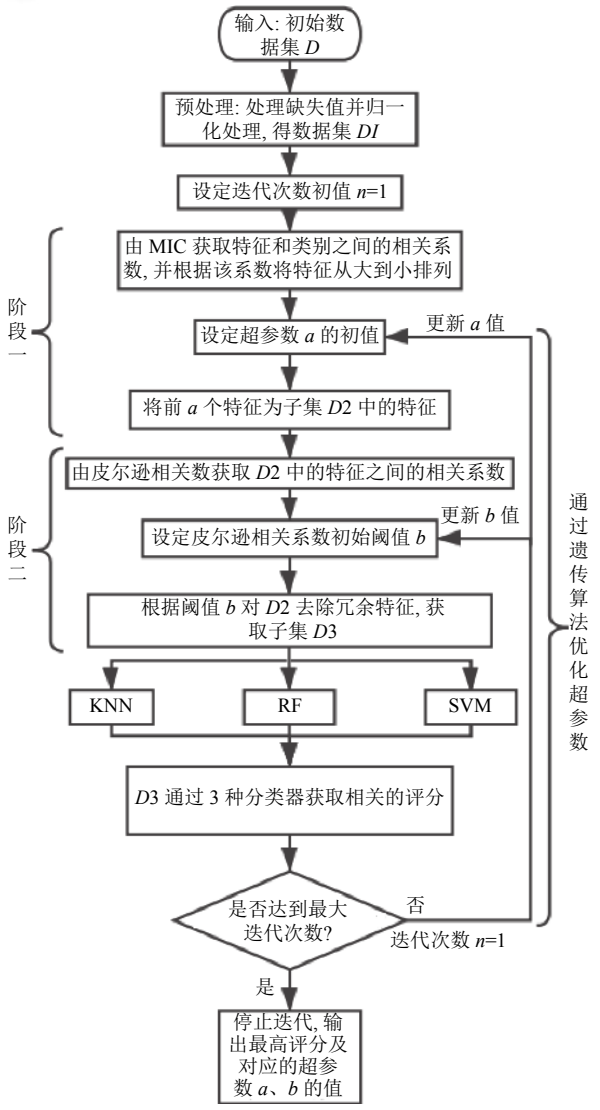

图 1 整体算法流程图

算法 1. 基于最大互信息系数的特征选择算法

输入: 数据集 $D1$.

步骤 1. 设定超参数 a.

步骤 2. 根据式 (5) 计算每个特征 $x_{(i)}$ 与类别 y 的最大互信息系数 $I_{\max}(x_{(i)};y)$.

步骤 3. 将特征按照最大互信息系数的大小, 从大到小排列, 获取排序后的特征集$\{x_{(1)},x_{(2)},\cdots,x_{(N)}\}$.

步骤 4. 从排序后的特征集中选择前 a 个特征, 构成特征子集 $D2=\{x_{(1)},x_{(2)},\cdots,x_{(a)}\}$.

输出: 特征子集 $D2$.

算法 2. 基于皮尔逊相关系数的特征选择算法

输入: 特征子集 $D2$.

步骤 1. 设定皮尔逊相关系数冗余阈值 b.

步骤 2. 根据式 (6) 计算 $D2$ 中所有特征之间的皮尔逊相关系数.

步骤 3. 将皮尔逊相关系数大于阈值 b 的两个特征中最大互信息值较小的特征删除.

步骤 4. 反复执行步骤 3 直至特征子集中所有特征的皮尔逊相关系数值都大于 b, 该特征子集记作 $D3$.

输出: 最终特征子集 $D3$.

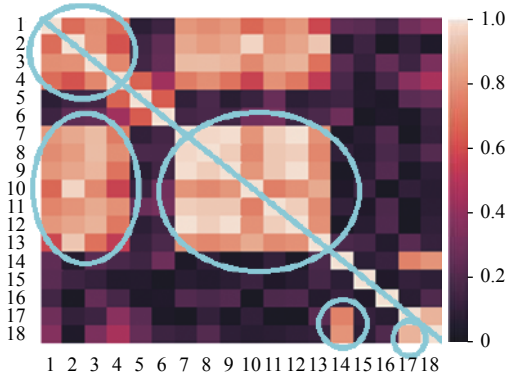

图 2 Vehicle 数据的皮尔逊相关系数热点图

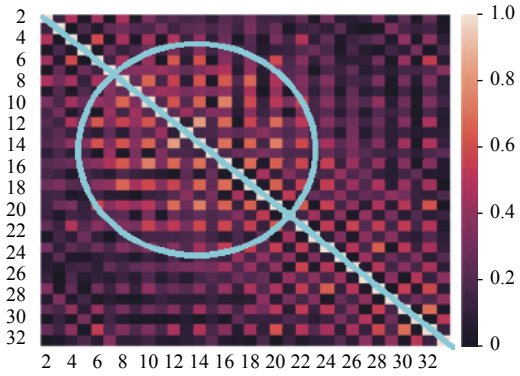

图 3 Ionosphere 数据的皮尔逊相关系数热点图

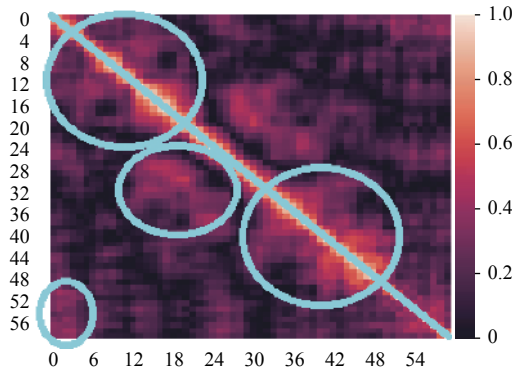

图 4 Sonar 数据的皮尔逊相关系数热点图

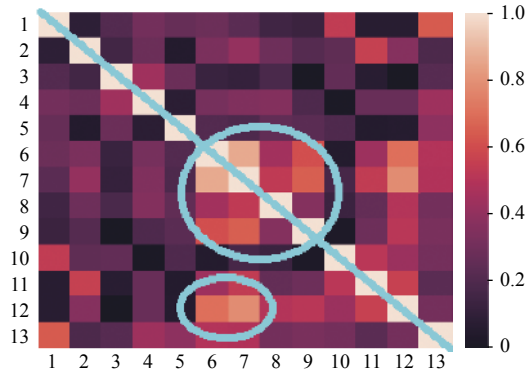

图 5 Wine 数据的皮尔逊相关系数热点图

2.3 使用遗传算法实现超参数的自动优化

本文采用遗传算法自动优化 3.1 和 3.2 节中的超参数 a 和 b. 遗传算法优化超参数的步骤描述如算法 3 所示.

此外染色体个数 $NIND$ 设为 5. 最大迭代次数 $MAXGEN$ 设为 20. 子代和父代个体不相同的概率 $GGAP$ 设为 0.9. 遗传算法选择方式 $SELECTSTYLE$ 选用轮盘赌选择法 rws. 基因变异的概率 PM 由源代码中的 0.1 修改为由式 (7) 获取, 令 PM 随着迭代次数 t 的增加而不断减小, 使得算法在搜索前期扩大搜索空间, 不容易陷入局部最优解, 而且在搜索后期也不会因为 PM 过大导致子集状态有较大突变.

$$PM=0.9-\frac{(t-1)\times 0.9}{MAXGEN-1} \tag{7}$$

综上所述, 本文提出的遗传算法可对前两步骤中的待优化超参数 a、b 自动进行优化, 并以适应度函数 $Fitness$ 作为目标函数, 将 $Fitness$ 取得最小值时对应的 a、b 作为最优参数.

算法 3. 使用遗传算法实现超参数的自动优化

步骤 1. 初始化染色体个数 *NIND* 和各个染色体中的超参数 *a* 和 *b*.

步骤 2. 根据超参数 *a*、*b* 结合阶段一、二的步骤获取各个染色体对应的特征子集, 通过适应度函数fitness = $AR + B\frac{M}{N}$计算各个染色体的适应度值 *fitness*, 并记录本轮获取的最小适应度值 *fitness**和对应的参数 *a**、*b**(其中 *R* 表示给定分类器的平均错误率, *M* 表示所选子集中特征的个数, *N* 表示初始数据集中所有特征的个数, 其中 *A* 和 *B* 分别是来自文献[18]与分类准确率和子集长度对应的两个参数, $A \in [1,0]$, $B \in [0,1-A]$).

步骤 3. 设定迭代次数 *t* 的初值和最大迭代次数 *MAXGEN*.

步骤 4. 使用遗传算法更新超参数 *a* 和 *b*, 迭代次数$t = t + 1$, 重复步骤 2, 若本轮最优适应度值 *fitness**小于历史最优适应度值 *fitnessbest*, 则$fitness_{best} = fitness*$, 并覆盖对应的超参数 *a**和 *b**, 其中超参数 $a \in [2, N]$, $b \in [Min(PCCs), Max(PCCs)]$, *Min*(*PCCs*) 和 *Max*(*PCCs*) 分别表示各个子集的特征之间的皮尔逊相关系数的最小值和最大值. 其中 *a* 的阈值设置为[2, *N*]是为了确保搜索算法的搜索空间完整, 最小值取 2 而不是取 1 则是考虑到代码参数个数需求.

步骤 5. 若迭代次数 *t* 等于最大迭代次数 *MAXGEN*, 终止迭代并输出历史最优适应度值 *fitnessbest* 和对应超参数 *a*、*b*.

3 实验分析

为检测上文提出的 MICP-GA 算法的可靠性, 本文选取了 5 个 UCI 标准数据集. 表 2 和表 3 分别列出了这些测试数据集的具体信息以及直接将各数据集初始数据用于不同分类器时的分类准确率.

表 2 测试数据集

数据集	样本数	特征数	标签数
Glass	214	9	7
Breast	699	9	2
Sonar	208	60	2
Ionosphere	315	34	2
Vehicle	846	18	7
Wine	178	13	3

在测试中, 使用 Python3 进行编程, 并且使用常规的十折交叉验证用于检验. 接下来使用 3 种特征选择方法对上述数据集进行测试和比较, 3 种方法分别是 MICP-GA、单独用 MIC 特征选择以及单独使用皮尔逊相关系数特征选择. 在测试过程中, 将 3 种特征选择方法与三种典型分类器, K 最近邻分类器 (K-Nearest Neighbors, KNN)、随机森林分类器 (Random Forest, RF) 和支持向量机 (Support Vector Machine, SVM) 结合后, 对上述数据集进行测试. 结果显示平均最高准确率如表 4 所示, 平均特征选取率如表 5 所示, 平均适应度函数如表 6 所示. 对比表 3 和表 4 的分类准确率, 可见本文提出的算法对原始数据的分类准确率有明显提升, 说明 MICP-GA 在搜索最优特征选择的组合时能够有效跳出局部最优, 确保搜索到的解是近似全局最优解, 同时说明 MICP-GA 充分考虑特征和类别之间的包括线性和非线性的关系, 解决了传统相关系数处理非线性关系表现不好的问题, 并根据 MIC 评分来判断删除每对冗余特征中的哪一个, 较好地结合了运用 MIC 和 PCCs 的特征选择从而使得分类效果比单独使用 MIC 或 PCCs 时更优. 同时, 结合表 5 中基于不同特征选择方法的分类器的平均特征选取率, 也表明了该算法在降低数据维度方面有较优的表现, 能够有效剔除和类别不相关的特征以及冗余特征, 为后续学习算法节省大量的运算成本, 提升了运行效率. 综上所述, 本文提出的算法充分考虑了特征和类别之间的线性相关性以及特征之间的冗余性, 能够有效对数据进行降维同时保证数据集的分类准确率保持不变甚至提升分类准确率.

表 3 基于不同分类器的初始数据集的平均分类准确率

数据集	KNN	RF	SVM
Glass	0.652	0.696	0.658
Breast	0.959	0.963	0.949
Sonar	0.799	0.771	0.839
Ionosphere	0.817	0.829	0.822
Vehicle	0.678	0.725	0.784
Wine	0.951	0.958	0.958

表 4 基于不同特征选择方法的分类器的平均最高准确率

数据集	KNN			RF			SVM		
	MICP-GA	MIC	PCCs	MICP-GA	MIC	PCCs	MICP-GA	MIC	PCCs
Glass	**0.706**	0.706	0.673	**0.858**	0.858	0.718	**0.864**	0.864	0.680
Breast	**0.974**	0.974	0.969	**0.980**	0.980	0.966	**0.983**	0.983	0.963
Sonar	**0.871**	0.843	0.814	**0.949**	0.925	0.794	**0.892**	0.875	0.866
Ionosphere	**0.899**	0.887	0.833	**0.887**	0.863	0.857	**0.890**	0.876	0.854
Vehicle	**0.733**	0.730	0.705	**0.823**	0.803	0.787	**0.837**	0.833	0.822
Wine	**0.985**	0.985	0.963	**0.990**	0.990	0.974	**0.986**	0.986	0.977

表 5 基于不同特征选择方法的分类器的平均特征选取率

数据集	KNN			RF			SVM		
	MICP-GA	MIC	PCCs	MICP-GA	MIC	PCCs	MICP-GA	MIC	PCCs
Glass	**0.444**	0.556	0.889	**0.444**	0.556	0.889	**0.444**	0.556	0.889
Breast	**0.444**	0.556	0.889	**0.444**	0.444	0.889	**0.444**	0.444	0.889
Sonar	**0.567**	0.700	0.867	**0.567**	0.733	0.833	**0.750**	0.850	0.900
Ionosphere	**0.412**	0.471	0.824	**0.412**	0.500	0.765	**0.441**	0.559	0.765
Vehicle	**0.278**	0.500	0.667	**0.222**	0.444	0.556	**0.222**	0.389	0.556
Wine	**0.539**	0.539	0.846	**0.462**	0.462	0.846	**0.462**	0.462	0.846

表 6 基于不同特征选择方法的分类器的平均适应度值

数据集	KNN			RF			SVM		
	MICP-GA	MIC	PCCs	MICP-GA	MIC	PCCs	MICP-GA	MIC	PCCs
Glass	**0.294**	0.294	0.327	**0.142**	0.142	0.282	**0.136**	0.136	0.320
Breast	**0.026**	0.026	0.031	**0.020**	0.020	0.034	**0.017**	0.017	0.037
Sonar	**0.129**	0.157	0.186	**0.051**	0.075	0.206	**0.108**	0.155	0.134
Ionosphere	**0.101**	0.113	0.167	**0.113**	0.137	0.143	**0.110**	0.124	0.146
Vehicle	**0.267**	0.270	0.295	**0.177**	0.197	0.263	**0.163**	0.167	0.178
Wine	**0.015**	0.015	0.037	**0.010**	0.010	0.026	**0.014**	0.014	0.023

4 结论与展望

特征选择算法的优劣对模型的预测准确率有着重要影响, Filter 可以快速高效地去除冗余特征及不相关特征, 但是无法保证获取的特征子集准确率达标. Wrapper 更倾向于获取准确率更高的特征子集, 所以其时间成本一般远高于前者. 本文针对以上两种方法的特点, 将其进行结合提出了 MICP-GA 方法, 所提算法兼顾了两者的优点并且在 UCI 标准数据集中也有较好的表现. 但是因为结合了遗传算法以及最大互信息系数, 导致时间复杂度比传统特征选择方法更高, 根据文献[10]可见传统的全局搜索算法在性能和运算复杂度上并没有较大差别, 所以没有将 GA 替换成其他全局搜索算法. 而如果使用单解的搜索算法如模拟退火 (Simulated Annealing, SA) 算法替换 GA 确实可以较快获取解, 但鉴于 SA 极其容易陷入局部最优, 因此不考虑用 SA 搜索最优特征子集. 所以如何加快搜索速度是后续进一步的研究课题.

参考文献

1 Cai J, Luo JW, Wang SL, *et al*. Feature selection in machine learning: A new perspective. Neurocomputing, 2018, 300: 70–79. [doi: 10.1016/j.neucom.2017.11.077]

2 Yang JM, Liu YN, Liu Z, *et al*. A new feature selection algorithm based on binomial hypothesis testing for spam filtering. Knowledge-Based Systems, 2011, 24(6): 904–914. [doi: 10.1016/j.knosys.2011.04.006]

3 Wang YY, Peng WX, Qiu CH, *et al*. Fractional-order darwinian PSO-based feature selection for media-adventitia border detection in intravascular ultrasound images. Ultrasonics, 2019, 92: 1–7. [doi: 10.1016/j.ultras.2018.06.012]

4 Zhu XF, Li XL, Zhang SC, *et al*. Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(6): 1263–1275. [doi: 10.1109/TNNLS.2016.2521602]

5 Chang XJ, Yang Y. Semisupervised feature analysis by mining correlations among multiple tasks. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2294–2305. [doi: 10.1109/TNNLS.2016.2582746]

6 Zhu PF, Xu Q, Hu QH, *et al*. Multi-label feature selection with missing labels. Pattern Recognition, 2018, 74: 488–502. [doi: 10.1016/j.patcog.2017.09.036]

7 Yan XA, Jia MP. Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and mRMR feature selection. Knowledge-Based Systems, 2019, 163: 450–471. [doi: 10.1016/j.knosys.2018.09.004]

8 Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Boston, MA: Springer, 1998: 7–8.

9 Hancer E, Xue B, Zhang MJ. Differential evolution for filter feature selection based on information theory and feature ranking. Knowledge-Based Systems, 2018, 140: 103–119. [doi: 10.1016/j.knosys.2017.10.028]

10 Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Applied Soft Computing, 2018, 62: 441–453. [doi: 10.1016/j.asoc.2017.11.006]

11 Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research, 2003, 3(6): 1157–1182.

12 Lai C, Reinders MJT, Wessels L. Random subspace method for multivariate feature selection. Pattern Recognition Letters, 2006, 27(10): 1067–1076. [doi: 10.1016/j.patrec.2005.12.018]

13 Boussaïd I, Lepagnot J, Siarry P. A survey on optimization metaheuristics. Information Sciences, 2013, 237: 82–117. [doi: 10.1016/j.ins.2013.02.041]

14 Abusamra H. A comparative study of feature selection and classification methods for gene expression data of glioma. Procedia Computer Science, 2013, 23: 5–14. [doi: 10.1016/j.procs.2013.10.003]

15 Mafarja MM, Mirjalili S. Hybrid whale optimization algorithm with simulated annealing for feature selection. Neurocomputing, 2017, 260: 302–312. [doi: 10.1016/j.neucom.2017.04.053]

16 Ma CW, Ma YG. Shannon information entropy in heavy-ion collisions. Progress in Particle and Nuclear Physics, 2018, 99: 120–158. [doi: 10.1016/j.ppnp.2018.01.002]

17 Reshef DN, Reshef YA, Finucane HK, *et al*. Detecting novel associations in large data sets. Science, 2011, 334(6062): 1518–1524. [doi: 10.1126/science.1205438]

18 Ghamisi P, Benediktsson JA. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. IEEE Geoscience and Remote Sensing Letters, 2015, 12(2): 309–313. [doi: 10.1109/LGRS.2014.2337320]