计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(7):174−179 [doi: 10.15888/j.cnki.csa.006965]

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

融合序列后向选择与支持向量机的混合式特征选择算法①

吴清寿, 刘长勇, 林丽惠

(武夷学院 数学与计算机学院, 武夷山 354300)
(认知计算与智能信息处理福建省高校重点实验室, 武夷山 354300)
通讯作者: 吴清寿, E-mail: wyswuqsh@163.com

摘　要: 维度灾难是机器学习任务中的常见问题, 特征选择算法能够从原始数据集中选取出最优特征子集, 降低特征维度.提出一种混合式特征选择算法, 首先用卡方检验和过滤式方法选择重要特征子集并进行标准化缩放, 再用序列后向选择算法 (SBS) 与支持向量机 (SVM) 包裹的 SBS-SVM 算法选择最优特征子集, 实现分类性能最大化并有效降低特征数量.实验中, 将包裹阶段的 SBS-SVM 与其他两种算法在 3 个经典数据集上进行测试, 结果表明, SBS-SVM 算法在分类性能和泛化能力方面均具有较好的表现.

关键词: 混合式特征选择; 序列后向选择; 支持向量机; 降维

引用格式: 吴清寿,刘长勇,林丽惠.融合序列后向选择与支持向量机的混合式特征选择算法.计算机系统应用,2019,28(7):174–179. http://www.c-s-a.org.cn/1003-3254/6965.html

Hybrid Feature Selection Algorithm for Fusion Sequence Backward Selection and Support Vector Machine

WU Qing-Shou, LIU Chang-Yong, LIN Li-Hui

(School of Mathematics and Computer Science, Wuyi University, Wuyishan 354300,China)
(Fujian Provincial Key Laboratory of Cognitive Computing and Intelligent Information Processing, Wuyishan 354300, China)

Abstract: Dimensional disaster is a common problem in machine learning tasks. The feature selection algorithm can select the optimal feature subset from the original data set and reduce the feature dimension. A hybrid feature selection algorithm is proposed. Firstly, the chi-square test and filtering method are used to select the important feature subsets and normalize scale, and then SBS-SVM wrapped by SBS and SVM. The algorithm selects the optimal feature subset to maximize the classification performance and effectively reduce the number of features. In the experiment, the SBS-SVM in the parcel stage and the other two algorithms are tested on three classical data sets. The results show that the SBS-SVM algorithm has better performance in classification performance and generalization ability.

Key words: hybrid feature selection; sequential backward selection; Support Vector Machine (SVM); dimension reduction

① 基金项目: 福建省自然科学基金 (2019J01835, 2017J01651, 2017J01780); 福建省中青年教师教育科研项目 (JAT170608); 认知计算与智能信息处理福建省高校重点实验室开放课题 (KLCCIIP2017104)
Foundation item: Natural Science Foundation of Fujian Province (2019J01835, 2017J01651, 2017J01780); Mid-aged and Young Teachers Program for Education Research of Fujian Province (JAT170608); Open Fund of Fujian Provincial Key Laboratory of Cognitive Computing and Intelligent Information Processing (KLCCIIP2017104)
收稿时间: 2018-10-26; 修改时间: 2018-11-19, 2019-01-07; 采用时间: 2019-01-31; csa 在线出版时间: 2019-07-01

引言

在现实应用场景中, 样本的属性维度经常成千上万, 而机器学习的各类方法中, 经常面对的问题之一是距离的计算, 当维度过大的时候, 内积的计算都难以实现. 高维数据中的冗余特征增加了计算的复杂度, 也增加了机器学习的难度. 特征降维是缓解维度灾难的重要途径, 其主要包括特征提取和特征选择[1]. 特征提取通过将原始特征映射到低维空间, 并希望能在低维空间中保持原始数据的相关信息; 而特征选择方法是在评价准则的指导下, 从数据集的全体特征中选取出满足评价指标的部分特征 (也称为最优特征子集) 的过程. 根据特征子集的评估方法, 可以将特征选择分为过滤式、封装式和嵌入式三种模型[2].

按照特征选择算法中的搜索策略分类, 可将其分为穷举式、启发式与随机式[3]. Focus 算法[4]是典型的穷举法, 其能够识别出所有的最优特征子集. 序列前向选择与序列后向选择及其各类改进算法[5]属于启发式方法, RFR[6]是一种基于序列前向选择的自底向上的特征选择方法, 属于有监督学习算法. 随机式无需遍历所有组合, 具有更好的时间复杂度, 文献[7]对粒子群优化和蚁群优化算法在特征选择上的应用进行了总结.

将支持向量机 (Support Vector Machines, SVM) 与特征选择结合的研究中, SVM-RFE[8]是基于 SVM 的回归特征消去方法, 使用 RFE (Recursive Feature Elimination, 递归特征消除) 在特征排序过程选取出最优特征子集. K-SVM-RFE[9]将 SVM-RFE 和特征聚类算法结合, 用于去除无关基因. Tan 等人[10]将 SVM 加入遗传算法, 并将其与特征选择算法进行封装, 基于分类准确度迭代求解. 文献[11]提出基于特征子集区分度衡量准则, 以 SVM 为分类工具, 通过计算特征子集对分类的联合贡献来考虑特征的相关度.

文献[12]提出一种基于混合式特征选择算法, 主要对特征过滤阶段的 CFS 算法进行改进, 未对过滤后的精简子集优化处理以提高模型的训练速度. 文献[13]提出的两级特征选择方法对原始特征集合中的噪声和无关特征进行过滤, 通过 Fisher 和信息增益同时计算, 最后进行交叉选择精简子集, 该方法在特定场景下可以有效提高分类准确率, 但增加了过滤阶段的时间开支.

本文提出一种混合式特征选择算法, 先用卡方检验计算特征与类别的相关度, 再采用过滤式选择方法得到相关度高的重要特征子集, 并对重要特征子集进行标准化缩放, 之后在包裹了 SVM 的序列后向选择算法 (SBS-SVM) 上进一步选出最优特征子集. 为了评估算法的有效性, 本文将 SBS-SVM 算法与 SBS-NB 和 SBS-KNN 的实验结果进行比较, 对不同算法选取出的最优特征子集组合数量、最高准确度, 基于最优特征子集训练的训练集与数据集的准确度与拟合度、维度缩减率等进行了讨论.

1 混合式特征选择算法模型

过滤式特征选择方法可以根据特征与目标的相关性对特征进行过滤, 其主要方法之一单变量特征选择 (univariate feature selection) 分别计算变量的相关统计指标, 并根据统计结果筛选出重要特征子集. 本文选择卡方检验对特征与类别的相关性进行检验.

特征缩放 (feature scaling) 可以提升分类算法和优化算法的性能, 在进行包裹式特征选择之前, 需要将重要特征子集进行特征缩放. 特征缩放的主要方法有归一化和标准化, 其中, 标准化方法使得特征值呈正态分布, 利于训练阶段的权重更新, 同时, 标准化方法还可以保持异常值的信息, 进一步减少异常值对算法的影响. 标准化的方法如式 (1):

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x} \tag{1}$$

其中, μ_x和σ_x分别表示数据集某个特征列的均值和标准差.

包裹式特征选择方法将学习器的分类准确度作为最优特征子集的评价标准, 其输入是上一步骤得到的重要特征子集. 在优选后的重要特征子集中, 采用序列后向选择可以较为快速的获得最优特征子集. 考虑到误差的评分与所采用的分类器及其精确度计算方式相关, 而支持向量机在高维特征空间中的模式识别优势明显, 故将支持向量机与序列后向选择方法进行包裹封装.

本文提出的混合式特征选择算法主要流程如图 1 所示.

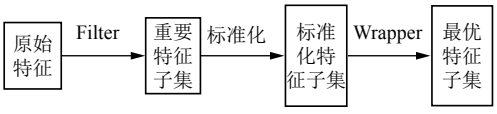

图 1 混合式特征选择算法流程

2 算法实现

2.1 序列后向选择算法

序列后向选择算法 (Sequential Backward Selection,

SBS) 是序列特征选择算法的一种典型应用, 其基于贪婪搜索算法, 用于将原始的n维特征空间缩减到一个k维特征子空间, 其中$k < n$.

SBS 算法的基本思想是: 在分类性能衰减最小的约束下, 通过逐步移除不相关的特征, 选取出与问题最相关的特征子集, 从而提高分类学习算法的计算效率, 通过缩减不相关特征可以降低模型的泛化误差, 提高模型的预测能力.

2.2 支持向量机

支持向量机是一种分类方法, 其基本模型是定义在特征空间上的间隔最大的线性分类器. SVM 通过确定一个分离超平面, 使得不同类别的样本分别处于超平面的两侧, 并使得两个不同类别的支持向量到超平面的距离之和最大, 其中, 支持向量到超平面的距离也称为间隔 (或几何间隔). 间隔越大, 则算法所产生的超平面的分类结果越鲁棒, 其泛化能力也越强.

对于训练集$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, 其中$x_i \in X = R^n$, $y_i \in \{-1, +1\}$, $i = 1.2, \cdots, N$, 其分离超平面为$w^T x + b = 0$.

假设x_i是支持向量, 则其他样本点x_j到分离超平面的函数间隔$\hat{\gamma}_j$必然大于或等于支持向量到分离超平面的函数间隔$\hat{\gamma}$. SVM 的基本思路是求解一个几何间隔最大而又能正确分离样本点的分离超平面, 其可表示为一个约束最优化问题:

$$\min_{w,b} \frac{1}{2}\|w\|^2, \ \text{s.t.} \ y_j(w^{\mathrm{T}} x_j + b) - 1 \geqslant 0 \tag{2}$$

在多数数据集上会存在噪音 (离群点), 这将造成线性不可分的问题, 其本质就是样本点(x_j, y_j)无法满足函数间隔$\hat{\gamma}_j \geqslant 1$的约束条件, 给分离超平面的构建增加了难度, 甚至无法构建超平面. 解决问题的方法是通过对每个样本点(x_j, y_j)添加一个松弛变量$\xi_j \geqslant 0$, 使得$\hat{\gamma}_j + \xi_j \geqslant 1$, 并设置一个惩罚因子$C > 0$, C是常数, 其值越大对误分类的惩罚越大, 对噪声的容忍度越低. 松弛变量的引入将增强模型的容错能力. 加入松弛变量后的目标函数和约束条件为:

$$\begin{cases} \min\limits_{w,b,\xi} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{j=1}^{N}\xi_j \\ \text{s.t.} \ y_j(w^{\mathrm{T}} x_j + b) + \xi_j - 1 \geqslant 0, \ \xi_j \geqslant 0, i = 1.2, \cdots, N \end{cases} \tag{3}$$

序列最小优化 (Sequential Minimal Optimization, SMO) 算法[14]将一个复杂的优化问题转化为一个比较简单的两变量优化问题, 运用 SMO 可以快速求解α^*和b. 由此可得:

分离超平面为

$$\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_i) + b = 0 \tag{4}$$

分类决策函数为

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_i) + b\right) \tag{5}$$

2.3 混合式特征选择算法

输入: 数据集 Ds
特征集 X
分类算法 SMO
停止条件控制参数k
输出: 最优特征子集 X^-

算法步骤:
1. 用卡方检验方法计算特征与类别的相关性Rf
2. 根据Rf选择 n 个重要特征集合X_n
3. 对X_n进行标准化, 得到标准化特征子集SX_n
4. 用 SBS-SVM 对SX_n进行特征优选, 得到最优特征子集X^-

SBS-SVM 算法中, 从SXn中逐一减少特征数, 并用 SMO 算法根据当前特征集进行分类, 评估其分类性能, 最后得到一个或多个最优特征子集, 并在实验阶段讨论各最优特征子集的实际分类性能.

SBS-SVM 算法

```
1.  d=|SXn|
2.  X⁻=SXn
3.  err_d=∞
4.  while d>k
5.    for c in comb(X⁻,d−1)
6.      res=SMO(Ds,c)
7.      err*=1−Accuracy(res)
8.      if (err*<err_d) then
9.        err_d=err*
10.       X_d⁻=c
11.     end if
12.   end for
13.   X⁻=X_d⁻
14.   d=d−1
15. end while
```

算法第 5–12 行对候选最优特征子集X^-按照指定的特征子集数量进行排列组合, 并对每一个特征子集组合求误差 (性能损失).

3 实验

实验的数据集选择 Wine, Iris 和 Breast Cancer Wisconsin (WDBC) 等 3 个 UCI 的常用数据集. Wine 数据集通过化学成分分析推断出葡萄酒的起源, 其包含 178 个样本, 每个样本有 13 个特征值, 特征值的数据类型是整数或实数, 都是连续变量, 所有样本分为 3 类. Iris 数据集通过鸢尾花的 4 个特征预测花卉的种类, 共有 3 种类型, 样本数为 150 个. WDBC 是有关乳腺癌的数据集, 包含 569 个样本, 30 个用于诊断的特征, 诊断结果为 2 种类型.

为了对比算法的效果, 本文将 SVM、KNN (K Nearest Neighbor, K 最近邻) 和 NB (Naive Bayes, 朴素贝叶斯)3 种分类算法封装到序列后向选择算法中, 对算法在全体特征、最优特征子集上的分类准确率、拟合程度等进行比较. 实验中, SBS-SVM 算法的核函数采用线性核函数, SBS-KNN 算法中 K 的值取 3, SBS-NB 算法的先验概率为正态分布. 数据集中 70% 的样本划分为训练集, 30% 的样本划分为测试集.

本文的实验环境: Intel(R) Core(TM) i5-6500 CPU 3.20 GHz, 8 GB 内存, Windows7 操作系统, 算法采用 Python3.6 实现.

3.1 评价指标

本研究采用的评价指标有三个: 分类准确率、F1 得分和维度缩减率.

分类准确率的定义如式 (6):

$$accuracy(y,y') = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(y_i' = y_i) \quad (6)$$

其中, y_i'是第i个样本的预测值, y_i是真实值, $1(x)$是指示函数, 当$y_i' = y_i$时, 即x值为真, 则$1(x)$的值为 1, 否则为 0.

F1 得分的定义如式 (7):

$$F1 = \frac{2\times P\times R}{P+R} = \frac{2\times TP}{2\times TP+FP+FN} \quad (7)$$

其中, P是查准率, R是召回率, TP是真正例, FP是假正例, FN是假反例. 对于多分类问题的 F1 得分计算, 本研究采用 micro-F1, 其定义如式 (8):

$$micorF1 = \frac{2\times microP\times microR}{microP+microR} \quad (8)$$

维度缩减率的定义如式 (9):

$$Dr = 1-(SF/AF) \quad (9)$$

其中, SF是最优特征子集中的特征数, AF是全体特征数. Dr的取值范围为[0, 1], 其值越大, 表示维度缩减的效果越好.

3.2 特征过滤实验

特征过滤实验中, 先用卡方检验方法计算各特征与类别的相关度, 再选择相关度较高的特征构成重要特征子集. 如何确定重要特征子集的数量是一个需要考虑的问题, 表 2 中, 缩减率最小的值为 0.385, 根据这个结果, 考虑将重要特征子集的数量设置为全体特征数的 70%. 表 1 为按 70% 优选后的重要特征子集.

表 1 卡方检验优选后的重要特征子集

数据集	重要特征子集	特征数
Iris	1, 3, 4	3
Wine	1, 2, 4, 5, 6, 7, 9, 10, 12, 13	10
WDBC	1, 2, 3, 4, 6, 7, 8, 11, 13, 14, 16, 17, 21, 22, 23, 24, 25, 26, 27, 28, 29	21

3.3 算法在重要特征子集上的实验

为了较为全面的评价 SBS-SVM 算法效果, 本文对三个数据集的重要特征子集进行排列组合, 并将其与 SBS-KNN 和 SBS-NB 两种分类算法进行对比测试, 结果如图 2, 图 3, 图 4 所示.

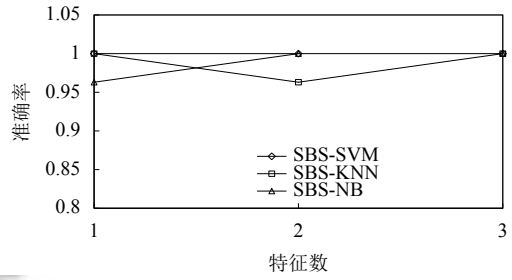

图 2 在 Iris 数据集上的最高准确率

图 2 中, SBS-SVM 算法在特征数为 1、2、3 的三种情况下都能获得最大准确率为 1 的效果, 而 SBS-KNN 和 SBS-NB 两种算法都只能在两类特征组合中取得 1 的准确率. 因为 Iris 数据集的特征数较少, 三种算法的效果对比并不明显.

在 Wine 数据集上的实验结果中, SBS-SVM 共有 5 种最优特征子集的组合可以获得 1 的准确率, SBS-KNN 算法可以获得 3 种组合的最高准确率, 而 SBS-NB 算法最高只能获得 0.968 的准确率. 实验结果如图 3 所示.

在 WDBC 数据集上的实验结果中, SBS-SVM 共有 6 种最优特征子集的组合可以获得 1 的准确率,

SBS-NB 算法可以获得 4 种组合的最高准确率 1, 而 SBS-KNN 算法只在 1 种组合上得到 1 的准确率. 实验结果如图 4 所示.

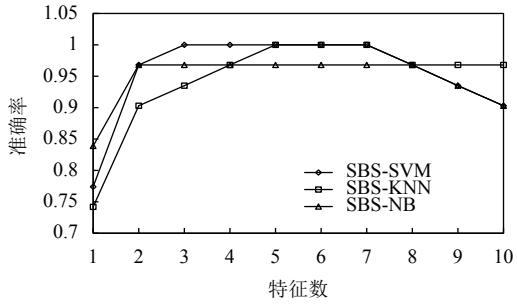

图 3 在 Wine 特征子集组合上的最高准确率

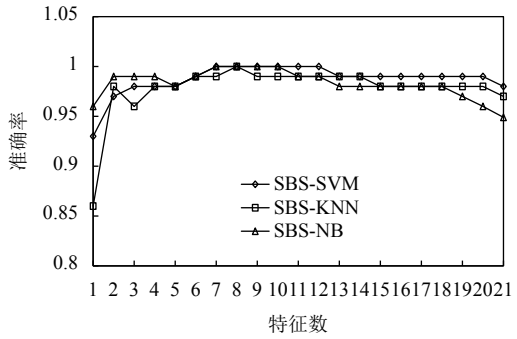

图 4 在 WDBC 特征子集组合上的最高准确率

从以上实验结果对比可以看成, SBS-SVM 能够在各种数据集及其特征子集组合上取得较高的分类准确率, 尤其是特征数较多的数据集, 其效果更加明显.

3.4 算法在最优特征子集上的实验

为了进一步验证基于 SBS-SVM 算法的有效性, 将上一节中识别出的最优特征子集组合进行进一步的测试. 实验通过将最优特征子集用于重新训练原训练集, 并用原测试集进行测试, 模型评估采用 F1 得分, 实验结果如图 5、图 6 所示.

图 5 中, 在特征数为 8 的特征子集组合上, 算法在训练集上和测试集上都能获得较高的 F1 得分, 其值分别为 0.952 和 0.946, 且两者之间的差距最小, 意味着该特征子集组合具有较好的泛化能力.

图 6 中, 在特征数为 12 的特征子集组合上, 算法在训练集上和测试集上的 F1 得分分别为 0.980 和 0.971.

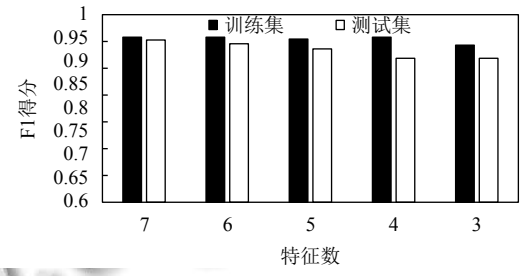

图 5 在 Wine 最优特征子集上的实验

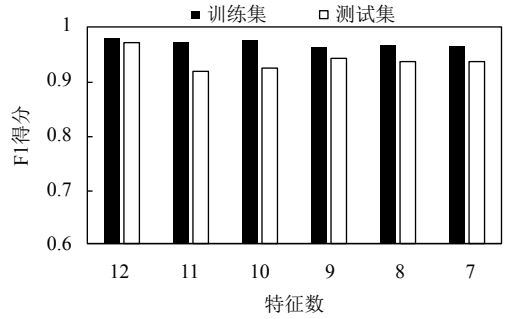

图 6 在 WDBC 最优特征子集上的实验

表 2 中, 将三种算法在 WDBC 和 Wine 两个数据集上进行测试, 选取在训练集和测试集上 F1 得分较高且两个值较为接近的最优特征子集进行比较, 并计算其维度缩减率 *Dr*. 可以看出, SBS-SVM 算法能够在训练集合测试集上都取得较高的 F1 得分, 且两者之间的差较小, 避免了过拟合的问题. 在 WDBC 数据集上, SBS-SVM 与 SBS-KNN 的 *Dr* 值略小于 SBS-NB, 在 Wine 数据集上, SBS-SVM 与 SBS-KNN 的 *Dr* 值相同, 略高于 SBS-NB. 从维度缩减率看, 两个数据集的最优特征子集在不同算法中的表现差异较小.

表 2 综合拟合度及 F1 得分选择的特征数与*Dr*值对比

数据集	SBS-SVM				SBS-KNN				SBS-NB			
	训练集	测试集	SF	Dr	训练集	测试集	SF	Dr	训练集	测试集	SF	Dr
WDBC	0.980	0.971	12	0.667	0.975	0.942	12	0.7	0.965	0.959	11	0.800
Wine	0.952	0.946	7	0.385	0.963	0.965	7	0.462	0.954	0.933	8	0.385

3.5 SBS-SVM 中松弛变量的影响实验

SBS-SVM 算法中, 松弛变量的选择对实验结果也会有一定的影响, 本部分实验在两个数据集上对松弛变量的变化与 F1 得分的关系进行讨论. 在 Wine 数据

集中选择的特征数为 7, 在 WDBC 数据集中选择的特征数为 12.

如图 7 所示, 松弛变量的取值对训练集和测试集都有一定的影响, 但总体影响不大, 在 Wine 数据集中, 训练集 F1 得分最大值和最小值的差为 0.01, 测试集的 F1 得分最大值和最小值的差为 0.0175, 在 WDBC 数据集上的差别也很小, 在训练集和测试集上的 F1 得分最大值和最小值之差分别为 0.008 和 0.019. 所以, 本研究中所提出的特征选择算法对松弛变量不敏感.

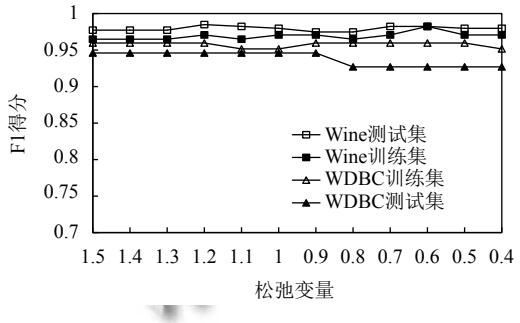

图 7 松弛变量实验

4 结束语

本文将特征选择分为两个步骤, 首先对特征进行过滤式选择, 得到重要特征子集; 之后, 对数据进行标准化缩放; 在包裹式特征选择阶段, 将序列后向选择算法与支持向量机进行封装, 提出一种 SBS-SVM 特征选择算法, 将其与 SBS-KNN 和 SBS-NB 两种算法在 Wine 等 3 个数据集上进行实验, 比较了算法所选择的不同特征子集的最高准确率, 并根据特征子集在训练集和测试集上的 F1 得分和拟合度进行最优特征子集选择. 实验结果表明, SBS-SVM 算法在 F1 得分和拟合度上都具有较好的效果, 但其所选择的最优特征子集在维度缩减率指标上与其他两种算法区分度不明显. 在今后的研究中, 针对无标签数据普遍存在的问题, 进行无监督的特征降维将是值得关注的研究领域.

参考文献

1 黄铉. 特征降维技术的研究与进展. 计算机科学, 2018, 45(6A): 16–21, 53.

2 Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4): 491–502. [doi: 10.1109/TKDE.2005.66]

3 初蓓, 李占山, 张梦林, 等. 基于森林优化特征选择算法的改进研究. 软件学报, 2018, 29(9): 2547–2558. [doi: 10.13328/j.cnki.jos.005395]

4 Almuallim H, Dietterich TG. Learning Boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 1994, 69(1-2): 279–305. [doi: 10.1016/0004-3702(94)90084-1]

5 Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognition Letters, 1994, 15(11): 1119–1125. [doi: 10.1016/0167-8655(94)90127-9]

6 Fujarewicz K, Wiench M. Selecting differentially expressed genes for colon tumor classification. International Journal of Applied Mathematics and Computer Science, 2003, 13(3): 327–335.

7 Kabir MM, Shahjahan M, Murase K. A new hybrid ant colony optimization algorithm for feature selection. Expert Systems with Applications, 2012, 39(3): 3747–3763. [doi: 10.1016/j.eswa.2011.09.073]

8 Mao Y, Zhou XB, Xia Z, *et al*. A survey for study of feature selection algorithms. Pattern Recognition and Artificial Intelligence, 2007, 20(2): 211–218.

9 叶小泉, 吴云峰. 基于支持向量机递归特征消除和特征聚类的致癌基因选择方法. 厦门大学学报 (自然科学版), 2018, 57(5): 702–707.

10 Tan KC, Teoh EJ, Yu Q, *et al*. A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications, 2009, 36(4): 8616–8630. [doi: 10.1016/j.eswa.2008.10.013]

11 谢娟英, 谢维信. 基于特征子集区分度与支持向量机的特征选择算法. 计算机学报, 2014, 37(8): 1704–1718.

12 雷海锐, 高秀峰, 刘辉. 基于机器学习的混合式特征选择算法. 电子测量技术, 2018, 41(16): 42–46.

13 武小年, 彭小金, 杨宇洋, 等. 入侵检测中基于 SVM 的两级特征选择方法. 通信学报, 2015, 36(4): 2015127.

14 Platt JC. Fast training of support vector machines using sequential minimal optimization. Schölkopf B, Burges CJC, Smola AJ. Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press, 1998: 185–208.