直接验证的封装式特征选择方法

汪文勇[1]，刘 川[1]，赵 强[1]，沈晓明[2]，丘晓彤[3]

(1. 电子科技大学计算机科学与工程学院 成都 611731；2. 国网浙江省电力公司电力科学研究院 杭州 310014；
3. 电子科技大学格拉斯哥学院 成都 611731)

【摘要】封装式特征选择算法可以准确地选择出有价值的特征，但是其评价过程伴随着极大的时间复杂度。为此，该文针对封装式特征选择算法中时间复杂度最高的交叉验证评价环节，提出了可以替代交叉验证的特征集直接评价方法——LW测量。进一步，将该方法与封装式特征选择算法中常用的序列搜索策略相结合，提出了改进的序列前(后)向搜索特征选择算法SFS-LW(SBS-LW)。通过在2个UCI数据集上与传统的基于交叉验证的封装式特征选择算法进行3组对比实验，结果表明该改进特征选择方法具有与传统方法近似的分类精度，但在时间复杂度上则有数倍的改善。

关 键 词 特征选择； 序列搜索算法； 分类； 时间复杂度； 封装式方法
中图分类号 TP391.4 **文献标志码** A doi:10.3969/j.issn.1001-0548.2016.04.013

An Improved Wrapper Method for Feature Selection

WANG Wen-yong[1], LIU Chuan[1], ZHAO Qiang[1], SHEN Xiao-ming[2], and QIU Xiao-tong[3]

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731;
2. Zhejiang Electric Power Research Institute Hangzhou 310014;
3. UoG-UESTC Joint School, University of Electronic Science and Technology of China Chengdu 611731)

Abstract The wrapper feature selection methods can achieve high classification accuracy，however, its cross-validation scheme in evaluation phase is very expensive in terms of computing resource consumption. In this paper, we propose a new statistical LW-measure which can replace the cross-validation scheme to evaluate feature sets. Furthermore, two improved wrapper algorithms, i.e. sequential forward selection-LW (SFS-LW) and sequential backward selection-LW (SBS-LW), are presented for feature selection, on the basis of combination of LW-measure and sequence search algorithms. Three groups of experiments conducted on two University of California, Irvine (UCI) datasets show that the proposed algorithms can not only obtain the similar classification accuracy to that of the traditional wrapper methods, but also are nearly ten times faster than the traditional ones.

Key words feature selection; sequence search algorithm; text classification; time complexity; wrapper methods;

特征选择是模式识别和机器学习领域的核心问题和热点研究方向之一[1]。随着信息技术的发展以及互联网规模和应用领域的不断扩大，生物信息分析、金融数据挖掘、互联网海量文本、图片信息处理等众多研究领域的数据分析需求增多，数据特征域的规模极速增长，给学习算法带来“维度灾难”问题[2]。在分类问题中，不同的特征区分对象的类别和状态的能力是不同的，重要的特征区分能力强，与类别标签相关性高[3]。与之相反，冗余的特征不仅会影响分类算法的性能，同时还会带来额外的计算开销。特征选择通过排除不相关的和冗余的特征实现数据降维，是从原始特征域中选出最优特征子集的过程。特征选择通过选出具有代表性的特征子集，提高了算法效率，减少了计算开销，同时避免了过拟合问题，提高了泛化能力[4-5]。

正是由于特征选择为数据分析和数据理解带来很多益处，因此受到了研究者们的关注，并提出了许多特征选择方法。通常特征选择方法可以分为：过滤式(filter)[3,6-7]，封装式(wrapper)[7-9]和嵌入式(embedded)[10]。

过滤式方法通过某种准则对所有特征进行评分，通过分值排序来判断特征的重要程度。通常采用的准则包括相关性测量，类内和类间距离[11]，以及信息熵等。常用的方法包括信息增益(information

收稿日期：2016－05－15
基金项目：教育部-中国移动科研基金(MCM20130661)；计算机网络及应用四川省工程实验室基金(20160001)
作者简介：汪文勇(1967－)，男，教授，主要从事网络测量及性能管理、无线传感器网络等方面的研究.

gain, IG)[7]，互信息(mutual information, MI)[12]，卡方统计(chi-square, CS)[13]，交叉熵(cross entropy, CE)[14]和T-test[15]等。过滤式方法时间复杂度低，可以快速缩小特征集规模，但是所选特征数量难以确定，而且过滤式方法只关注单独的特征，忽略了特征之间的组合性能。

封装式方法最显著的特征是需要结合分类算法。封装式方法首先通过搜索策略在特征集上选出候选特征子集，然后分类算法作为引导算法对特征子集进行评价，迭代地进行这一过程，直到选出符合要求的特征子集[16]。因此，封装式方法可以达到比过滤式方法更高的精确度[17]。但是，封装式方法的时间复杂度远高于过滤式方法。

嵌入式方法将特征选择过程与算法学习过程结合起来，特征选择与学习过程同步进行，典型的学习算法包括ID3，C4.5等，利用决策树递归生成过程来进行特征选择。嵌入式方法比封装式方法时间复杂度低，比过滤式方法高，但是精确度没有封装式方法高且鲁棒性差。

对比三类特征选择算法，封装式方法在精确度上有优势，但是受限于时间复杂度过高。实际上，造成复杂度高的根本原因在于封装式方法需要结合分类算法对候选特征子集进行交叉验证评价[1,3]。虽然，交叉验证(cross-validation)[18-20]是对分类效果进行评价最普遍的方法，但是，反复的交叉验证带来了巨大的计算消耗[21-23]。在许多特征维度高的应用领域，如文本分类，基因分析[24]等，计算消耗会达到难以接受的程度，使得封装式方法难以被广泛应用。

为了优化封装式方法的执行效率，需要一种更加直接的评价方法来代替交叉验证，在特征搜索过程中，高效地评价候选特征子集。在交叉验证中，候选特征子集被划分为训练集和测试集，训练集被用来训练分类模型并把该模型应用到测试集上。通过某种测量(比如 F_1 [25])可计算出测试集的真实类别划分与基于分类模型所得的划分之间的差异。实际上，该候选特征子集真实的类别划分也可以被看成是基于某种聚类模型所得的划分，因此，可以采用聚类算法中的内部评价方法直接对该候选特征子集进行测量。当然，这样的测量方法必须具有以下特征：1) 精确度高，能识别出不相关的和冗余的特征，也就是说所选出的特征子集应用到分类算法上可以实现较高的分类精度；2) 时间复杂度低，减少计算消耗是改进封装式方法的初衷；3) 抗干扰性强，增加噪声样本点，不会带来该测量的跃变。目前，满足以上要求的特征集评价测量是没有的。

因此，本文提出了一种新的特征集评价测量方法(LW)，并把该方法与序列搜索策略相结合，提出了改进的封装式特征选择方法。当特征子集中的类别间隔距离大时，LW会有较高的值，说明类别线性可分程度高。反之，说明类别线性可分程度低。此外，LW拥有线性时间复杂度，因此，在封装式方法中引入LW，可以极大的减少特征子集评价过程中交叉验证所造成的计算开销，同时还可以保证良好的分类精确度。

1 相关工作

封装式特征选择算法一般包含三个部分[26]：搜索策略，评价函数和验证函数。搜索策略用于搜索特征空间，产生候选特征子集。一般使用的策略有：穷举搜索，启发式搜索和随机搜索[27]。穷举搜索遍历所有可能的特征子集，一定可以发现最优特征子集，但是这已被证明是NP难问题[28]。即使有分支定界(branch and bound)[1]这类改进方法，但是依然会带来巨大的计算开销。启发式搜索方法主要指序列搜索，序列搜索依照某个方向遍历特征空间[17]，经典的序列搜索算法包括序列前向搜索(sequential forward selection, SFS)和序列后向搜索(sequential backward selection, SBS)[1,29]。随机搜索方法随机产生特征子集，如：遗传算法(GA)，蚁群算法(ACO)等。评价函数用于评价候选特征子集，在迭代过程中作为每一步的指导，而验证函数用于验证最终的分类性能。

由于封装式特征选择算法采用相同的评价函数和验证函数，因此可以实现较高的分类准确度。一般而言，评价候选特征子集可结合特定分类器，采用固定测试集或交叉验证的方式。固定测试集的好坏直接影响到整个特征选择的性能，因此，可靠性差。而交叉验证伴随着巨大的时间复杂度，导致算法效率低。

为了提高封装式特征选择算法的性能，一些研究者尝试将各种统计机器学习方法应用到封装式特征选择方法中。如朴素贝叶斯(Naïve Bayes)[30]、K最近邻(K-nearest neighbor, KNN)[16]、神经网络(neural network)[31]、决策树(decision tree)[32]、支持向量机(support vector machines, SVM)[5,9,33]等。由于分类算法本身特性的不同，封装式方法使用这些算法引导时，特征选择效率会表现出一些差异。但是正

如之前所说，造成封装式方法时间复杂度高的最大原因是反复训练分类器的交叉验证评价方法。因此，单纯改变机器学习算法并不能解决这个问题。

此外，一些研究者致力于搜索算法的改进。如模拟生物演化现象的一些随机搜索策略：遗传算法(genetic algorithm, GA)[34]，蝙蝠算法(bat algorithm, BA)[35]，蚁群算法(ant colony optimization, ACO)[36]等。这些随机搜索策略执行效率高、速度快，在一些领域取得了不错的成效，但是，由于其随机性，所以运行结果不确定。此外还包括引入序列搜索策略和浮动序列搜索策略的研究，其目的也是改进搜索候选特征子集的计算消耗。如，文献[33]提出了一种结合SVM和序列后向搜索的改进封装式模型。每轮迭代采用错误特征数目评估候选特征子集，据此对特征进行剔除。然而，在搜索策略上所做的改善，依然没有从根本上解决封装式方式时间复杂度高的问题。

除了以上改进分类算法和搜索策略，一些研究者提出了将过滤式和封装式相结合的方法。在过滤式方法的速度优势和封装式方法的性能优势上折中，采用混合式的方式进行特征选择。通常的做法是用过滤式方法做特征预选，缩减特征维度，然后执行封装式方法，从而期望达到高准确度、低计算消耗的目的。如，文献[5]设计了一种基于序列前向搜索和SVM的混合式特征选择方法(FS_SFS)。该方法总共有两个步骤：首先，使用一种新的指标，利用特征识别能力和相关性的过滤式方法；其次，执行SFS和SVM的封装式方法。

文献[16]采用混合式策略，设计了4种基于KNN的特征选择方法。首先是预选阶段，KKN结合SFS，KNN结合SBS，以及基于相关系数，依赖函数的两种过滤式方法，共计4种方法来评估候选特征子集；其次是封装式方法阶段，所有4种方式都采用KNN和穷举搜索策略来发现最优特征子集。

文献[31]同样使用混合式特征选择方法，其中过滤式阶段采用互信息(mutual information, MI)，封装式阶段采用神经网络。此外，文献[32]提出了一种利用随机森林的混合式特征选择方法。在过滤式阶段利用决策树对特征进行排序，在封装式阶段使用了序列搜索策略和交叉验证进行候选特征子集评估。一般而言，混合式方法在损失一部分特征选择精度的条件下提高了封装式方法的效率。

综上所述，尽管以上改进的封装式方法的效率得到了一定提高，但是并没有从本质上解决交叉验证的计算消耗问题。因此，本文提出了基于LW测量的封装式特征选择方法。

2　LW测量

支持向量机是基于统计学理论的机器学习方法，其主要思想是追求结构风险最小化，寻找使不同类别支持向量间隔最大的最优的超平面[37]。受支持向量最大间隔思想的启发，本文提出了一种类间间隔距离的度量方法。

定义 1　自由度(FD)对于向量空间中给定的聚类C_i和聚类C_j，则聚类C_i相对于聚类C_j的自由度定义为在聚类C_i与聚类C_j的边缘支持向量不发生重叠的条件下，聚类C_i往任意方向所能移动的最小距离，表示为$\mathrm{FD}_{i|j}$。

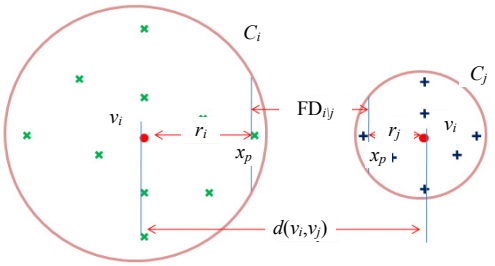


图1　聚类二维空间分布示例

显然，若自由度($\mathrm{FD}_{i|j}$或$\mathrm{FD}_{j|i}$)大于零，则聚类C_i和聚类C_j线性可分；反之，则聚类C_i和聚类C_j线性不可分。假设样本在向量空间中的分布是均匀的球体，如图1所示。为了保证线性时间复杂度，则两个聚类自由度的一种合理的计算方式为两个聚类质心之间的距离减去各自的类半径之和，如下所示：

$$\mathrm{FD}_{i|j}=\mathrm{FD}_{j|i}=\mathrm{d}(\boldsymbol{v}_i,\boldsymbol{v}_j)-(r_i+r_j) \tag{1}$$

式中，$\boldsymbol{v}_i$和$\boldsymbol{v}_j$分别为聚类C_i和聚类C_j的质心向量，采用余弦距离作为距离度量。通常，基于聚类C_*中的所有样本$\boldsymbol{x}_n\in C_*$则C_*的质心的计算为：

$$\boldsymbol{v}_*=\frac{1}{|C_*|}\sum_{\boldsymbol{x}_n\in C_*}\boldsymbol{x}_n \tag{2}$$

此外，r_i和r_j分别为聚类C_i和聚类C_j的类半径，其定义为该聚类中样本与其质心的最大距离。由于考虑到一个噪声点的加入可能带来该类半径发生数量级的变化，因此，本文中半径采用该聚类中样本与其质心的K_*个最大距离的平均值，有

$$r_*=\frac{1}{K_*}\sum_{k=1}^{K_*}d(\boldsymbol{x}_*^k,\boldsymbol{v}_*) \tag{3}$$

式中，参数K_*的取值为经验值。一般而言，K_*取值

为一个常数或者 $\mu|C_*|$，其中 μ 为一个经验分数。

对于多分类问题，设候选特征子集 $X=\{\boldsymbol{x}_n:n=1,2,\cdots,N;\boldsymbol{x}_n\in\mathbb{R}^D\}$ 为D维空间中的一组向量集合，且该向量集合所对应的标签集合为 $Y=\{y_n:n=1,2,\cdots,N;y_n\in\{1,2,\cdots,M\}\}$。对于每一个$D$维向量 $\boldsymbol{x}_n$ 有且有一个标签 y_n 把它标识到一个特定的聚类 $m\in\{1,2,\cdots,M\}$ 中。则候选特征子集X的LW测量定义为：

$$\mathrm{LW}_X=\frac{1}{M}\sum_{i=1}^{M}\min_{i\neq j,j=1,2,\cdots,M}\mathrm{FD}_{i|j} \tag{4}$$

LW是一种有标签特征集的评价测量方法，该方法旨在测量特征集的线性可分程度。LW越高，说明该特征集不同类别间的可分离程度越高，因此，该特征集适用于一个分类模型，特别是线性分类模型。

3 基于LW的封装式方法

实际上，LW作为一种直接的特征集评价测量方法，具有线性时间复杂度，执行效率高，可代替传统的交叉验证的评价方法。因此，本文结合LW与序列搜索算法提出了改进的封装式特征选择算法：SFS-LW和SBS-LW，具体算法如下：

SFS-LW算法流程：

1) 输入，原始特征集 F_e；2) 输出，最优特征集 F_o。

算法步骤：

1) 初始化目标特征集 F_o 为 $\varnothing$；

2) 按序加入 $f_c(f_c\in F_e)$ 特征形成候选特征子集 $F_{oc}=f_c\cup F_o$；

3) 利用式(4)计算候选特征子集 F_{oc} 的LW值并记录；

4) 重复步骤2)～3)，直到遍历所有特征；

5) 对记录的所有LW排序，选出值最高的候选特征子集，将对应 f_c 加入 F_o；

6) 重复步骤2)～5)直到满足终止条件或阈值。

SBS-LW算法流程与SFS-LW基本相同，区别在于SBS-LW初始化 F_o 为特征全集，通过删除特征形成候选特征子集 F_{oc}。

综上所述，SFS-LW和SBS-LW算法与传统封装式算法的主要区别就在于评价候选特征集的方式。显然，本文提出的两种改进算法相比于传统封装式算法在效率上将有很大提升。需要指出的是，LW还可以与其他搜索策略相结合。

4 实验与结果分析

本文基于UCI(University of California, Irvine)机器学习库的两个真实数据集Twenty Newsgroups和Gas对算法进行了实验。实验平台为Pentium(R) Dual-Core E6700 CPU 3.20 GHz，4 GB RAM。

4.1 数据集和评价指标

Twenty Newsgroups：是一个文本数据集，包括来自于20类新闻的20 000条消息，有4%转帖。其中的文章都是典型的帖子，因此其标题包括主题行，签名文件和引用其他文章的部分。由于考虑到传统封装式方法在高维数据集上计算开销太大，因此，在每个类别中随机抽样100个样本共计2 000个样本进行实验，分词后获取的特征维度为10 319。其次，本实验使用TF-IDF[38]作为特征词加权机制。

Gas Sensor Array Drift Dataset(Gas)[39]：该数据集是加利福尼亚大学(University of California，San Diego)科研机构于2008年1月～2011年2月(36个月)采集的实验数据。特征数据采集于6种不同气体中的传感器，包含了由128个实数特征描述的13 910个样本。该数据集可以应用于各项人工智能研究。

精确率(precision, P)和召回率(recall, R)是分类算法中最常用的评价指标。精确率(查准率)是指被正确判定属于某类别的样本数量与被判定属于该类别的全部样本数量的比值。召回率(查全率)是指被正确判别属于某类别的样本数量与实际属于该类别的样本数量的比值。由于精确率和召回率是相互的，单纯提高其中一个性能可能导致另一个性能的下降。为权衡精确率和召回率，本实验采用 F_1 度量作为分类性能指标，定义如下：

$$F_1=\frac{2RP}{P+R} \tag{5}$$

为了评价分类器在所有类别上的全局分类性能，通常采用微平均值(microF_1)和宏平均值(macroF_1)。本实验中对于传统的封装式特征选择算法，均采用LibSVM分类器，其采用径向基核函数，参数c设置为100，其余为默认值。评价方式为5阶交叉验证。对于LW测量，K_* 设置为 $0.1\times|C_*|$。

4.2 实验设计

本实验分为3组，每组之间的区别在于采用的搜索算法不同。第1组使用SFS搜索策略，第2组使用SBS搜索，而第3组采用了随机选择特征的策略。对于Gas数据集，每轮搜索迭代的特征数量为1。由于考虑到Twenty Newsgroups数据集的高维特性，每轮迭代的步长设定为100(约为总特征数量的1%)。

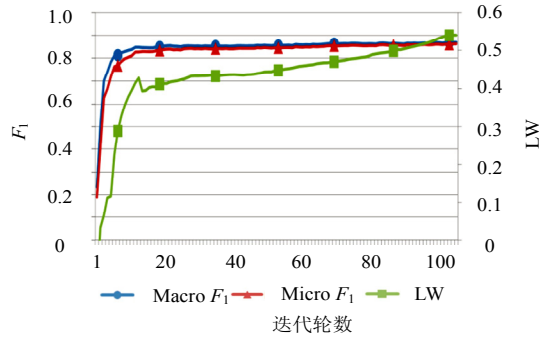
第1组实验使用SFS搜索，形成了两种特征选择算法：SFS-LW和SFS-SVM。首先，使用SFS-LW算法进行特征选择，根据每轮迭代选择的特征集，对其进行交叉验证并统计 F_1 测量，观察LW测量与 F_1 测量的变化趋势；其次，使用SFS-SVM算法进行特征选择，根据每轮迭代选择的特征集，对其进行LW测量统计，观察 F_1 与LW测量的变化趋势；

第2组实验使用SBS搜索策略，并分别采用了SBS-LW和SBS-SVM两种特征选择算法。其余设置与第一组实验相同。

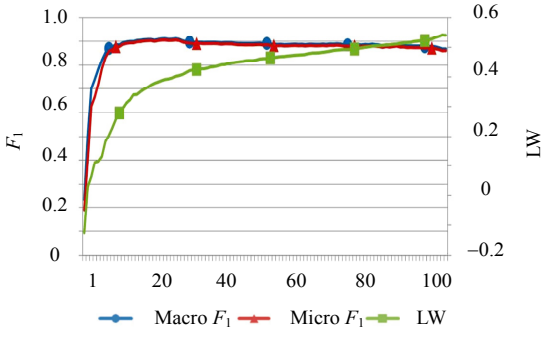
第3组实验采用了随机选择添加特征和随机选择删除特征两种策略。对每轮迭代中的特征子集统计其LW与 F_1 测量。该实验旨在观察LW和 F_1 测量在特征集变化下的变化趋势是否一致，并计算其相关性。由于该实验是随机选择特征，因此，相邻迭代过程中特征子集的评价指标也会有明显的大幅波动，此时是观察二者相关性的最好时机。

4.3　实验结果与分析

第1组实验在Newsgroups和Gas数据集上的结果分别如图2和图3所示。折线图中 Y 轴主坐标代表 microF_1 和 macroF_1 的值，副坐标表示LW的值，X 轴表示迭代轮数。图2a和图3a是由SFS-LW算法引导的特征选择实验；图2b和图3b由SFS-SVM算法的 microF_1 引导的特征选择实验。

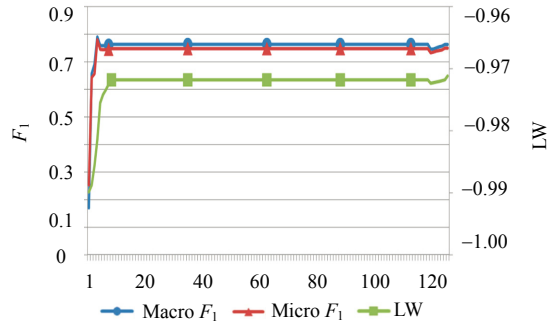


a. SFS-LW算法

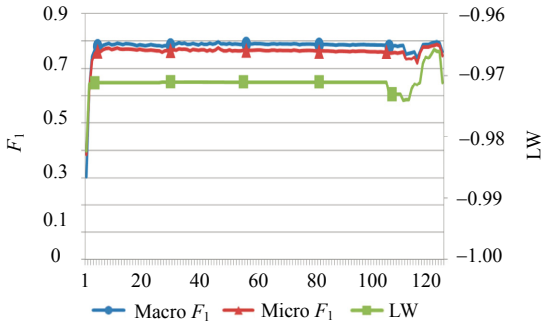


b. SFS-SVM算法

图2　SFS-LW算法和SFS-SVM算法在20Newsgroups数据集上的性能对比



a. SFS-LW算法



b. SFS-SVM算法

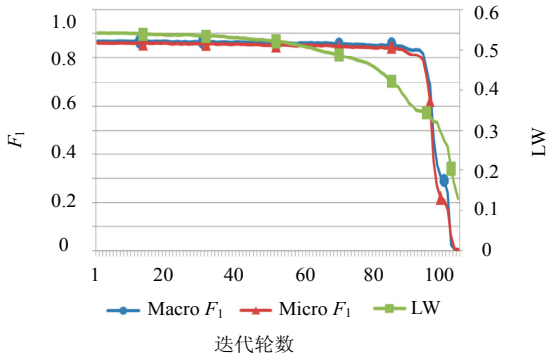
图3　SFS-LW算法和SFS-SVM算法在Gas数据集上的性能对比

在20Newsgroups数据集上，如图2a所示，从特征集为空开始，每次迭代增加1%的特征，F_1 与LW测量一开始极速增长，第10次迭代之后，F_1 测量趋于稳定并维持在0.80以上，然后增速开始放缓，最终在全部特征集时达到0.862的峰值；如图2b所示，类似地，F_1 测量先经历了一个快速的增长阶段，第28次迭代后 microF_1 达到峰值0.906，之后二者数值小幅下降并趋于平缓直至特征集增加到全集。

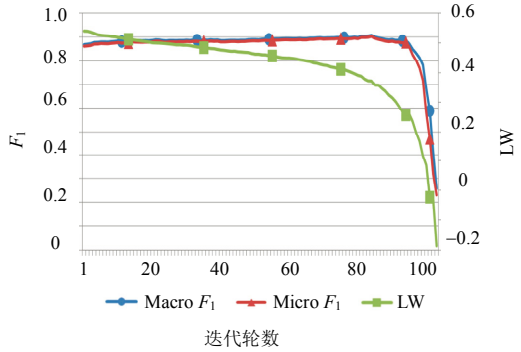
在Gas数据集上，如图3a所示，从特征集为空开始，每次迭代增加1个的特征，F_1 与LW测量在前4次迭代极速增长并达到峰值0.78，之后小幅下降并趋于稳定，microF_1 维持在0.74左右；如图3b所示，F_1 测量在前3次迭代达到峰值后趋于平缓，维持在0.75以上，在一些波动之后在第122次迭代后达到峰值0.789，随后小幅下降。

通过观察图2和图3不难发现以下结论：1) LW与 F_1 测量变化趋势基本保持一致，因此，说明两种评价测量方式有很强的相关性；2) 采用SFS-LW算法引导的特征选择实验(图2a和图3a)与采用SFS-SVM算法的 microF_1 引导的特征选择实验(图2b和图3b)都能使 F_1 测量很快达到峰值。虽然在LW测量引导下并没有选出可以使 microF_1 值更高的特征集，但是特征

的数量已得到大幅缩减并维持一个较高的分类精度，更重要的是拥有线性时间复杂度，相同环境下的时间消耗会在后面进行对比分析。



a. SBS-LW算法



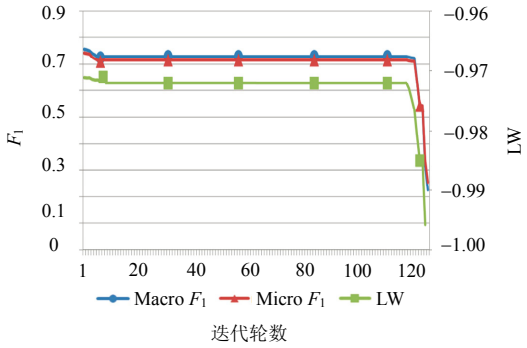
b. SBS-SVM算法

图4 SFS-LW算法和SFS-SVM算法在20Newsgroups数据集上的性能对比

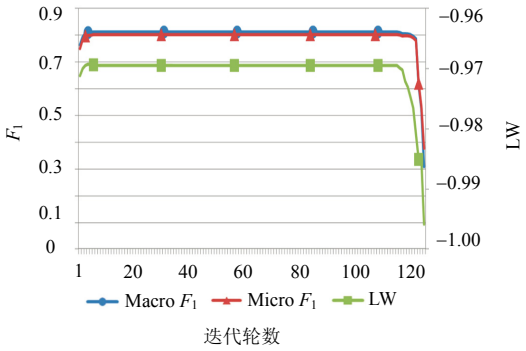
第2组实验的结果分别如图4和图5所示。图4a和图5a由SBS-LW算法引导的特征选择实验；图4b和图5b由SBS-SVM算法的microF_1引导的特征选择实验。

在20Newsgroups数据集上，如图4a所示，特征集从全集开始，虽然F_1测量缓慢下降，但microF_1基本维持在0.8以上，直到迭代95次之后才开始显著下降，这说明在特征选择过程中并没有剔除重要的特征；如图4b所示，特征集从全集开始，在特征剔除的过程中，F_1测量开始缓慢上升，在第85次迭代后microF_1达到峰值0.903，之后F_1性能开始显著下降。可以看出，两种特征选择算法虽然大幅减少了特征的数量，但是依然保持一个较高的分类效果，这符合特征选择的要求。

在Gas数据集上，如图5a所示，从特征集全集开始，每次迭代剔除1个特征，microF_1在前124次迭代都维持在0.70以上，之后开始明显下降直至特征集为空；如图5b所示，microF_1经过7次迭代达到峰值0.812，然后经历长时间平缓迭代后F_1测量才开始下降直至特征集为空。



a. SBS-LW算法



b. SBS-SVM算法

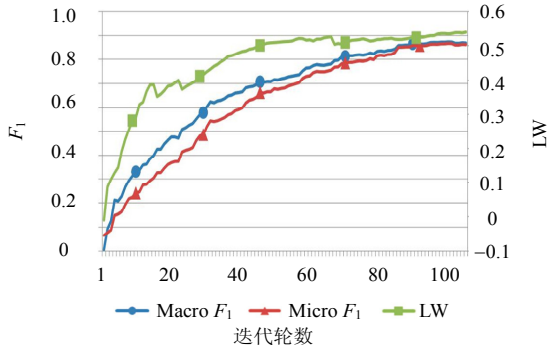
图5 SBS-LW算法和SBS-SVM算法在Gas数据集上的性能对比

此外，本实验同样可以得出在第1组实验中得出的结论。

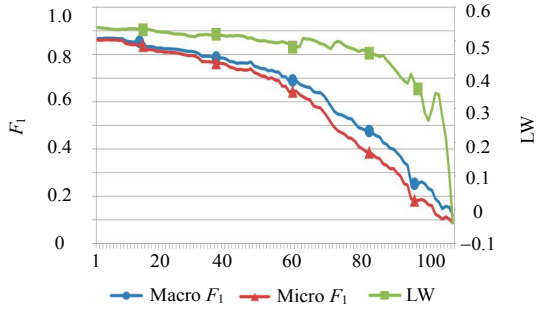
表1 各算法的时间消耗mins/轮

数据集	SFS-LW	SBS-LW	SFS-SVM	SBS-SVM
20Newsgroups	217.84	211.73	1 763.35	1 827.41
Gas	57.68	58.26	521.13	526.38

时间复杂度是衡量算法效率的重要指标，表1显示了基于LW的算法时间消耗远小于传统的交叉验证评价方法。在Twenty Newsgroups训练集上，SFS-LW算法平均迭代时间是217.84 min，SFS-SVM的平均迭代时间是1 763.35 min，这是前者的8.09倍。SBS-LW平均迭代时间是211.73 min，SBS-SVM的平均迭代时间是1 827.41 min，这是前者的8.63倍。在Gas训练集上，SFS-LW算法平均迭代时间是57.68 min。然而，SFS-SVM的平均迭代时间是521.13 min，这是前者的9.03倍。SBS-LW算法平均迭代时间是58.26 min，SBS-SVM的平均迭代时间是526.38 min，这是前者的9.04倍。



a. RFS算法

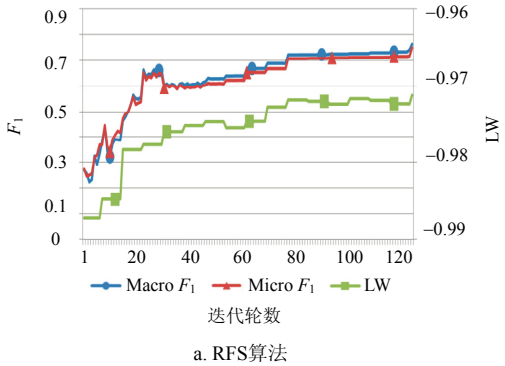


b. RBS算法

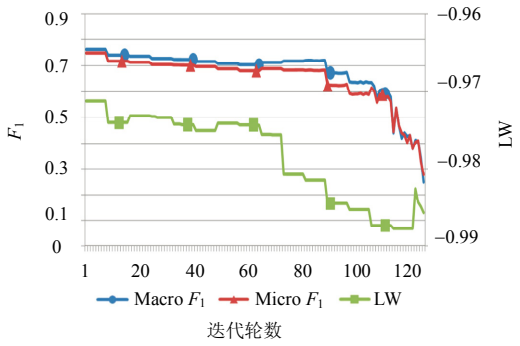
图6　随机特征选择在20Newsgroups数据集上的性能分析

第3组实验的结果分别如图6和图7所示。由于随机前向搜索(random forward search, RFS)和随机后向搜索(random backward search, RBS)并不会关注特征本身的特性，即不会关注特征本身的类别区分能力，因此，观测到的 F_1 和LW测量都呈现出明显的上下波动。显然，在此情况下，很容易观察出两种评价测量数值的相关性。

在Newsgroup数据集上，如图6a所示，随着特征的随机加入，F_1 和LW测量缓慢增加，并且呈现出类似的变化趋势；如图6b所示，随着特征的随机剔除，F_1 和LW测量缓慢减小，并且呈现出类似的变化趋势。此外，在Gas数据集中，可以观察到在Newsgroup数据集上类似的结论。两种方式下，F_1 和LW测量都具有明显一致的上升或下降情况。为了定量地分析两种评价测量数值的相关性，该组实验中两种测量的皮尔逊相关性统计如表2所示，该统计表明 F_1 和LW测量的相关性是非常高的。



a. RFS算法



b. RBS算法

图7　随机特征选择算法在Gas数据集上的性能分析

表2　第3组实验中LW与 F_1 测量的相关性统计

数据集	搜索算法	macroF_1 &LW	microF_1 &LW
20Newsgroups	RFS	0.957	0.926
	RBS	0.869	0.842
Gas	RFS	0.955	0.957
	RBS	0.778	0.821

此外，通过把第3组实验与第1组实验和第2组实验对比，发现在随机特征选择算法中明显没有第1组和第2组实验中的特征选择算法有效，这说明了LW和 F_1 测量作为封装式方法的引导评价方式都达到了特征选择的要求，也就是说无论是LW还是 F_1 测量作为特征选择的依据，都是可行的。

5　结 束 语

本文基于一个新的特征集评价方法LW和常用的序列搜索算法，提出了改进的封装式特征选择算法SFS-LW和SBS-LW。LW测量评价与交叉验证评价相比计算效率高，时间复杂度低，从根本上改善了封装式特征选择方法的应用瓶颈，最大程度地发挥封装式特征选择方法准确度高的优势。本文通过在真实数据集Twenty Newsgroups和Gas Sensor Array Drift Dataset上的一系列实验对其效果进行验证，其结果表明SFS-LW和SBS-LW算法可以取得和传统封装式方法相当的准确度，并节省大量时间。

参　考　文　献

[1] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. J Mach Learn Res, 2003, 3: 1157-1182.

[2] ABDI H, WILLIAMS. “Principal component analysis” Wiley interdisciplinary reviews[J]. Computational Statistics, 2010, 2: 433-459.

[3] KOHAVI R, JOHN G H. Wrappers for feature subset selection[J]. ArtifIntell, 1997, 97: 273-324.

[4] JUHA R. Overfitting in making comparisons between variable selection method[J]. Journal of Machine Learning Research, 2003, 3: 1371-1382.
[5] LIU Yi, ZHENG Yuan. FS_SFS: a novel feature selection method for support vector machines[J]. Pattern Recognit, 2006, 39: 1333-1345.
[6] LIU Huan, SETIONO R. A probabilistic approach to feature selection: a filter solution[C]//Proceedings of the Thirteenth International Conference on Machine Learning. Bari: [s.n.], 1996, 319-327.
[7] CHEN W, CHANG X, WANG H, et al. Automatic word clustering for text categorization using global information [C]//Asia Information Retrieval Symp. Beijing: Springer-Verlag, 2004, 1-11.
[8] XIONG M, FANG Z, ZHAO J. Biomarker identification by feature wrappers[J]. Genome Res, 2001, 11: 1878-1187.
[9] CHEN Gang, CHEN Jin. A novel wrapper method for feature selection and its applications[J]. Neurocomputing, 2015, 159: 219-226.
[10] PUDIL P, NOVOVICOVA N, KITTLER J. Floating search methods[J]. Pattern Recognition Letters, 1994, 15: 1119-1125.
[11] MICHAEL M, LIN W C. Experimental study of information measure and inter-intra class distance ratios on feature selection and orderings[J]. Systems Man & Cybernetics IEEE Transactions on, 1973, smc-3(2): 172-181
[12] LARKEY L S. Automatic essay grading using text categorization techniques[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne: ACM, 1998: 90-95.
[13] CAROPRESO M F, MATWIN S. SEBASTIANI F. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization[J]. Text Databases and Document Management: Theory and Practice, 2001, 5478: 78-102.
[14] MLADENIC D, GROBELNIK M. Feature selection for unbalanced class distribution and naive bayes[C]// Proceedings of the 16th International Conference on Machine Learning. [S.l.]: ICML, 1999: 258-267.
[15] CHAVES R, RAMÍREZ J. SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting[J]. Neuroscience Letters, 2009, 461(3): 293-297.
[16] UNCU O, TURKSEN I B. A novel feature selection approach: Combining feature wrappers and filters[J]. Information Sciences, 2007, 177(2): 449-466.
[17] SAEYS1 Y, INZA I. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 177(23): 2507-2517.
[18] JUANG B H, KATAGIRI S. Discriminative learning for minimum error classification[J]. IEEE Trans Signal Process, 1992, 40: 3043-3054.
[19] SEBBANA M, RICHARD N. A hybrid filter/wrapper approach of feature selection using information theory[J]. Pattern Recognition, 2002, 35: 835-846.
[20] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection[J]. Fourteenth International Joint Conference on Artificial Intelligence, 2001, 14: 1137-1143.
[21] GHEYAS I, SMITH L. Feature subset selection in large dimensionality domains[J]. Pattern Recognition, 2010, 43: 5-13.
[22] KUDO M, SKLANSKY J. Comparison of algorithms that select features for pattern classifiers[J]. Pattern Recognit, 2000, 33: 25-41.
[23] MAHESH P, GILES M. Feature selection for classification of hyperspectral data by SVM[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 45(5): 2297-2306.
[24] IN˜AKI I, PEDRO L. Filter versus wrapper gene selection approaches in DNA microarray domains[J]. Artificial Intelligence in Medicine, 2004, 31: 91-103.
[25] MARINA S, GUY L. A systematic analysis of performance measures for classification tasks[J]. Information Processing and Management, 2009, 45: 427-437.
[26] AHA D W, BANKERT R L. A comparative evaluation of sequential feature selection algorithms[C]//Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale: [s.n.] 1995, 112:1-7.
[27] DOAK J. Intrusion detection: the application of input selection, a comparison of algorithms and the application of a wide area network analyzer[D]. California: University of California, 1992.
[28] AMALDI E, KANN V. On the approximation of minimizing non zero variables or unsatisfied relations in linear systems[J]. Theoretical Computer Science, 1998, 209: 237-260.
[29] CARUANA R, SA V. Benefitting from the variables that variable selection discards[J]. JMLR, 2003, 3: 1245-1264.
[30] FERREIR A J, FIGUEIREDO A T. Incremental filter and wrapper approaches for feature discretization[J]. Neurocomputing, 2014, 123: 60-74.
[31] SHAMSUL H, ABDOLLAHIAN M. A hybrid wrapper–filter approach to detect the source(s) ofout-of-control signals in multivariate manufacturing process[J]. European Journal of Operational Research, 2014, 237: 857-870.
[32] CADENAS J M, GARRIDO M C, MARTÍNEZ R. Feature subset selection filter–wrapper based on low quality data[J]. Expert Systems with Applications, 2013, 40(16): 6241-625.
[33] SEBASTIN M, RICHARD W. A wrapper method for feature selection using support vector machines[J]. Information Sciences, 2009, 179: 2208-2217.
[34] DARYA C, ALEXANDRE S. Evolutionary ELM wrapper feature selection for Alzheimer's disease CAD on anatomical brain MRI[J]. Neurocomputing, 2014, 128: 73-80.
[35] DOUGLAS R, PEREIRA A M. A wrapper approach for feature selection based on Bat algorithm and optimum-path forest[J]. Expert Systems with Applications, 2014, 41:

2250-2258.
[36] TURKER T E, CUMHUR T, MERVE C. A wrapper-based approach for feature selection and classification of major depressive disorder–bipolar disorders[J]. Computers in Biology and Medicine , 2015, 64: 127-137.
[37] BURGES C J C. A tutorial on support vector machines for pattern recognition[J]. Dataing and Knowledge Discovery, 1998, 2(2): 121-167.
[38] NGUYEN T T, CHANG K, HUI S C. Supervised term weighting centroid-based classifiers for text categorization [J]. Knowledge and Information Systems, 2013, 35(1): 61-85.
[39] VERGARA A, VEMBU S, AYHAN T. Chemical gas sensor drift compensation using classifier ensembles[J]. Sensors and Actuators B: Chemical, 2012, 166: 320-329.

编 辑 蒋 晓

汪文勇(1967—)，教授，博士生导师。中国教育和科研计算机网(CERNET)专家委员会委员，中国下一代互联网(CNGI)专家委员会委员，下一代互联网核心网技术国家工程实验室(清华大学)技术委员会委员，下一代互联网关键技术和评测国家工程研究中心专家委员会委员，江苏省计算机网络技术重点实验室(东南大学)学术委员会委员，四川省计算机网络工程实验室学术委员会主任。主要研究方向为计算机网络。获国家及部省级科技进步奖共6次，发表论文40余篇，获国家发明专利9项。