特征选择方法综述

李郅琴,杜建强,聂　斌,熊旺平,黄灿奕,李　欢
江西中医药大学 计算机学院,南昌 330004

摘　要:特征选择作为一个数据预处理过程,在数据挖掘、模式识别和机器学习中有着重要地位。通过特征选择,可以降低问题的复杂度,提高学习算法的预测精度、鲁棒性和可解释性。介绍特征选择方法框架,重点描述生成特征子集、评价准则两个过程;根据特征选择和学习算法的不同结合方式对特征选择算法分类,并分析各种方法的优缺点;讨论现有特征选择算法存在的问题,提出一些研究难点和研究方向。

关键词:特征选择;搜索策略;评价准则;特征选择分类
文献标志码:A　**中图分类号:**TP301　doi:10.3778/j.issn.1002-8331.1909-0066

李郅琴,杜建强,聂斌,等.特征选择方法综述.计算机工程与应用,2019,55(24):10-19.
LI Zhiqin, DU Jianqiang, NIE Bin, et al. Summary of feature selection methods. Computer Engineering and Applications, 2019, 55(24):10-19.

Summary of Feature Selection Methods

LI Zhiqin, DU Jianqiang, NIE Bin, XIONG Wangping, HUANG Canyi, LI Huan
College of Computer Science, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China

Abstract: As a data preprocessing process, feature selection plays an important role in data mining, pattern recognition and machine learning. Through feature selection, the complexity of the problem can be reduced, and the prediction accuracy, robustness and interpretability of the learning algorithm can be improved. This paper introduces the framework of feature selection methods, and focuses on the two processes of generating feature subsets and evaluation criteria. The feature selection algorithms are classified according to different combinations of feature selection and learning algorithms, and the advantages and disadvantages of various methods are analyzed. The existing problems of existing feature selection algorithms are discussed, and some research difficulties and research directions are proposed.

Key words: feature selection; search strategy; evaluation criteria; feature selection classification

1　引言

人们早在20世纪60年代就开始研究特征选择,其涉及机器学习、模式识别等诸多领域[1-3]。起初,研究问题时希望获取尽可能多的特征,因为仅根据三两个特征建模效果不佳,可解释性不强,特征多可提供更多信息用以准确描述问题,研究结果将充满说服力。然而,随着研究对象难度增加,高维数据变得常见,例如使用高效液相和质谱联用仪获取的中医药物质基础实验数据,图像数据,以及各种基因表达数据。此类数据特征维度成千上万但样本量稀少,是典型的高维小样本数据集,易引发“维数灾难”“过拟合”,随着维度增加学习器性能可能不升反降等问题。正因为研究对象的复杂性,特征选择的发展日新月异。

高维数据的研究极具挑战。首先,高维数据会引起“维数灾难”,在保证学习算法预测精度的前提下,随着特征维度的提升,训练时样本需求量会呈指数形式增

基金项目:江西省科技厅重点研发计划(No.20171ACE50021);国家自然科学基金(No.61562045,No.61762051)。
作者简介:李郅琴(1996—),女,硕士研究生,研究领域为医药数据挖掘及机器学习;杜建强(1968—),通讯作者,男,博士,教授,CCF高级会员,研究领域为医药信息与数据挖掘,E-mail:jianqiang_du@163.com;聂斌(1972—),男,硕士,CCF会员,主要研究领域为中医药信息及数据挖掘;熊旺平,男,副教授,主要研究领域为医药数据挖掘、计算机体系结构;黄灿奕(1993—),男,硕士研究生,研究领域为医药数据挖掘及机器学习;李欢(1995—),女,硕士研究生,研究领域为医药数据挖掘及机器学习。
收稿日期:2019-09-05　**修回日期:**2019-10-21　**文章编号:**1002-8331(2019)24-0010-10
CNKI**网络出版:**2019-11-04, http://kns.cnki.net/kcms/detail/11.2127.TP.20191101.1122.002.html

长[4];其次,研究对象的复杂性以致生成数据集所需成本大,获取的样本少,所以有限样本甚至是小样本导致高维空间数据分布稀疏,高斯分布三法则不再适用;另外,有学者证明高维空间数据几乎分布于超球表面而非中心,超球中心作为高维空间的高密度区域却数据匮乏,这时一般的多元数据分析方法面对这种特殊的几何性质只会徒劳无益;最后,随着特征维数的增加,距离测度的作用被削弱,样本点之间的可区分性难以用样本点之间的距离衡量[5],所以高维空间中基于距离测度的算法效果甚微。

降维对高维小样本问题行之有效。维数约简是实现降维的关键技术,有特征提取和特征选择。

特征提取一般通过数学方法(如投影)将数据从高维特征空间映射到低维特征空间,典型的方法有主成分分析(PCA)[6]、核主成分分析(KPCA)[7]、典型相关分析(CCA)、Fisher线性判别分析(LDA)[8]、核线性判别分析(KDA)[9]、独立成分分析(ICA)[10]、核独立成分分析(KICA)[11]、多维尺度(Multi-Dimensional Scaling,MDS)变换[12]、奇异值分解方法(SVD)、偏最小二乘法判别分析(PLS-DA)等;神经网络亦与特征提取异曲同工,比如BP神经网络、卷积神经网络(CNN)等。经特征提取的新特征物理意义与原始特征相差甚远,甚至截然不同,提取到的特征可解释性弱,这在很多问题中难以接受。

特征选择使用某种评价准则从原始特征空间中选择特征子集,是一种数据预处理方式。迄今为止,学者们从特征子集能否识别目标、是否降低预测精度、会否改变原始数据类分布等多个角度对特征选择进行了定义[13-16]。总结来说,得到的特征子集要尽可能小,能够识别目标,可以解决问题,不能降低分类器或回归模型的预测精度甚至在一定程度上可以提高预测精度,并且不改变原始数据集的类分布,另外,特征子集的稳定性亦十分重要。

特征选择,顾名思义从原始特征空间中遴选"好的"特征,剔除"不好的"特征,"好的"特征指与任务相关的特征即相关特征,"不好的"特征指无关特征、冗余特征和噪声等。特征选择挑选的特征,物理意义一如既往,可解释性强,优势明显,所以近年来特征选择受到广泛关注,研究者众多,令其在各个领域大放异彩,包括中医药物质基础实验数据、基因表达数据、计算机视觉、生物信息学、目标识别等[17]。

根据特征和因变量的相关程度可分为无关、弱相关以及强相关特征;依据特征与特征的相关程度,分为冗余和非冗余特征。两者相结合,特征可以分为无关特征(无关特征分冗余非冗余没有意义)、弱相关且冗余、弱相关非冗余、强相关(数学定义上认为冗余特征必定是弱相关特征)四种。

理想的特征选择算法要实现去除无关、弱相关且冗余,且保留弱相关非冗余和强相关特征。而现有算法大都可以删除无关特征并且在一定程度上去冗余(如mRMR[18]、最小冗余最大分离[19]);也有一些算法冗余特征过度删除,导致丢失大量有用信息(如FCBF[20]),因此,亟需方法上的创新。

2 特征选择框架

传统的特征选择框架[16]如图1所示。

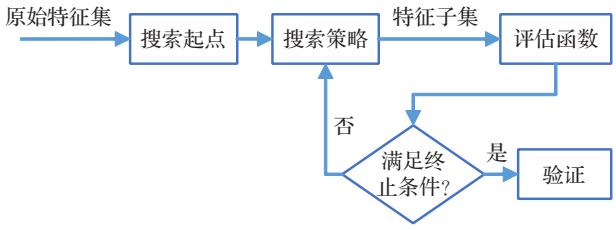

图1 传统的特征选择框架

整个框架包括了四个基本过程,即生成特征子集、评价特征子集、停止条件和验证结果。整个特征选择的过程是:首先使用空集(或全集)当作搜索起点,也即原始的已选特征子集;然后使用前向搜索策略从未选特征中选择一个特征加入到已选特征子集中(或使用后向搜索策略从已选特征子集中删除一个特征);已选特征子集每加入(或每删除)一个特征都需要进行评估;如果终止条件成立,则停止搜索并用学习算法验证其性能,否则继续使用前向搜索(或后向搜索)进行特征选择。

该框架是特征选择领域的基石,十有八九的传统特征选择算法皆源于该框架,生成特征子集和评价特征子集是四个过程的重中之重。

2.1 生成特征子集

生成特征子集过程有两大要领,一是搜索起点,所谓搜索起点,是指从何处开始遍历,随之对应何种搜索策略。二是搜索策略,即采用何种方式遍历原始特征集合以生成最优特征子集。

2.1.1 搜索起点

搜索起点决定了搜索方向,指出从何处开始遍历,四个不同的搜索起点,分别对应四个搜索策略:

(1)搜索起点为空集,每次加入一个得分最高(评价准则进行打分)特征到已选特征子集当中,这种搜索方式即为前向搜索。

(2)搜索起点是全集(原始特征子集),每次搜索,得分最低的特征将被删除,这种搜索方式是后向搜索。

(3)搜索起点前后方向双管齐下,搜索过程中,加入m个特征到已选特征子集当中,并且从其中删除n个特征,这种搜索方式称为双向搜索。

(4)搜索起点随机选择,搜索期间增加或删除特征亦采取随机的方式,叫随机搜索,它有机会使算法从局部最优中跳出来,一定几率获取近似最优解。

2.1.2 搜索策略

根据特征子集的搜索方式,可将搜索策略分成全局最优搜索、序列搜索和随机搜索[21]。序列搜索和随机搜索统称为启发式搜索策略,包括了四种搜索起点对应的四种搜索策略:前向、后向、双向和随机搜索。启发式搜索策略是最优化研究的一个重要分支,本质是在搜索过程中加入某些特性,使搜索有效的向最优方向前进,其效率远优于全局最优搜索,但一般只能得到局部最优解(次优解)。

(1)全局最优搜索策略

全局最优搜索,即找到原始特征集合的全局最优子集,迄今能够实现的只有穷举法和分支定界法[22]。该策略在特征维数不高的情况才有使用价值,可以想象,10 000维的原始特征集合特征子集高达$2^{10\,000}$个,使用穷举法不合实际,虽然分支定界法复杂度低于穷举法,但是面对高维数据也无能为力。随着维数的增加,该策略时间复杂度呈指数级增长,是一个NP-hard问题[23]。

(2)序列搜索

采用序列搜索的算法不计其数,序列搜索可分为前向搜索、后向搜索和双向搜索三大类。前向搜索中:序列前向搜索(SFS)每次贪心地把得分最高的特征加入到已选特征子集当中,序列前向浮动搜索(SFFS)和广义序列前向搜索(GSFS)等都是其改进策略;后向搜索的序列后向搜索(SBS)每次从已选特征子集中剔除一个特征,其改进策略有:序列后向浮动搜索(SBFS)、广义序列前向搜索(GSFS)、浮动广义后向搜索(FGSBS)等;双向搜索是前向搜索和后向搜索相辅相成的结合策略,既可增加特征,又可删减特征,有加q减r算法、广义加q减r算法等。

(3)随机搜索策略

随机搜索策略选取特征随机,不确定性强,本次和下次选择的特征子集千差万别,但通过启发式规则子集的变换幅度渐缓,逐渐接近最优特征子集,随机搜索有一定几率使算法跳出局部最优,即防止陷入局部最优,找到近似最优解,所以一般情况下随机搜索策略选取的特征子集会优于序列搜索。常用的随机搜索方法有模拟退火算法(Simulated Annealing,SA)[24]、差分进化(Differential Evolution,DE)、蚁群算法(Ant Colony Optimization,ACO)[25]、遗传算法(GA)[26]、量子进化算法(Quantum Evolutionary Algorithm,QEA)[27]、和声搜索算法(Harmony Search Algorithm,HSA)、粒子群算法(Particle Swarm Optimization,PSO,也称为鸟群算法)[28]、爬山搜索、人工免疫系统、禁忌搜索算法、Beam搜索、人工蜂群等。

(4)三种搜索策略的比较

三种搜索策略的优缺点见表1,全局最优搜索优势在于可实现全局最优,但时间复杂度太高,只在维度低时有使用价值;序列搜索的时间复杂度最低,但特征子集仅是局部最优的;随机搜索结合了前两种方法的优势,时间复杂度比全局最优低,比序列搜索高,同时可获得优于序列搜索的近似最优解。三种策略各有优缺点,因此实际应用中需"对症下药",选择合适的搜索策略。

表1　三种搜索策略的优缺点

搜索方式	解	效率
全局最优搜索	最优解	低下
序列搜索策略	局部最优解	高
随机搜索策略	近似最优解	较高

2.2 评价准则

作为特征评估方式,评价准则的优劣直接影响特征子集的优劣,即便同一算法,度量方式的差异亦可导致最优特征子集大相径庭。评价准则除了描述特征与类别(或被解释特征)的相关性,还能度量特征与特征的冗余性,因此,一种研究的趋势是提出不同的评价准则来改进算法,现有的评价准则可大体分为距离度量、一致性度量、依赖性度量、信息度量、分类正确率或分类误差五种[29],下面进行简要介绍。

2.2.1 距离度量标准

距离度量的核心是距离公式。特征和类别的距离越小,相关性越大;亦可根据特征与特征的距离大小判断两者是否冗余,距离越小,冗余性越大。距离度量分为几何距离和概率距离。几何距离有欧氏距离、明式距离、马氏距离等,例如算法Relief[30]、ReliefF[31]使用欧式距离度量特征与类别的相关性以及特征间的冗余性;概率距离使用概率定义类内距离和类间距离,"类内小,类间大"的特征与类别相关性大,常用的概率距离有Bhattacharyya、Kullback-Liebler、Kolmogorov、Chernoff、Matusita、Patrick-Fisher、Mahalanobis等。几何距离也适合回归问题,但概率距离只适合离散型特征,一般只用于分类问题。

2.2.2 一致性度量标准

一致性度量标准根据数据集中不一致样本数与样本总数的比值来衡量特征的重要性[32]。不一致因子[33]、Focus[34]、LVF[35]都是使用一致性度量标准的算法。该准则得到的特征子集规模较小,但只适合离散型特征,因此回归问题中无法使用,一般只在分类问题中应用,当然可以先行将连续特征离散化再使用该度量方式。

2.2.3 依赖性度量标准

依赖性度量使用统计原理评估特征与类别的相关性,分类和回归中都适用,例如f检验、t检验、Pearson相关系数、Fisher得分[36]等,这些统计相关系数只对特征与类别的相关性进行探讨,特征间的冗余性被忽略了,所以,学者们深入研究,文献[37]提出既考虑相关性,又兼顾冗余性的依赖性度量标准,文献[38]提出Constraint得分评价特征。

2.2.4 信息度量标准

不同于距离度量、依赖性度量只能描述线性关系,信息度量标准可以衡量特征间、特征与类别之间的非线性关系,所以信息度量一直是研究热点,多年来提出的信息度量标准比比皆是,有互信息[39]、信息增益、加入冗余惩罚的互信息[40-46]、最小描述长度、条件互信息[47]、归一化的互信息(见表2)等。

表2 归一化的互信息度量

归一化的互信息	度量公式
信息相关系数 ICC [46]	$ICC(X;Y)=\frac{I(X;Y)}{H(X,Y)}$
不确定系数 CU [41]	$CU(f,s)=\frac{I(f;s)}{H(s)}$
对称不确定性 SU [20]	$SU=\frac{2I(c;f)}{H(f)+H(c)}$
决策依赖相关性 Qc [44]	$Qc(f,s)=\frac{I(C;f)+I(C;s)-I(f,s;C)}{H(c)}$
NI [45]	$NI(f;s)=\frac{I(f;s)}{\min\{H(f),H(s)\}}$
SR [48]	$SR(c;s)=\frac{I(c;s)}{H(c,s)}$

以上信息度量标准只能处理离散型特征,2011年Reshef提出的最大信息系数(MIC)[49]可直接处理离散型特征,尤其适合回归问题中的特征选择。

2.2.5 分类正确率或分类错误率度量标准

分类正确率或分类错误率度量标准可衡量特征子集的整体性能。具体做法为:利用已选特征子集训练分类器,通过分类器的分类正确率或者分类错误率衡量特征子集整体性能的优劣。若是回归问题,那么使用已选特征子集建立回归模型,通过可决系数R^2、均方根误差等指标评价已选特征子集。

2.2.6 评价准则的比较

以上五种评价准则各有优缺点(见表3)。概括来说,前四种大多只估计单个特征得分,而分类正确率或分类错误率可评价子集性能,综合性能最好,但效率最差;距离、一致性和依赖性度量分类性能一般但效率较好;信息度量会优于距离、一致性和依赖性度量,分类性能好效率也高。但现在CPU性能好,对时间复杂度要求没那么严格,因此实际应用中以获得最优子集为目标,通常会将分类误差率或正确率度量和其他四种度量方式结合使用,先使用前四种度量方式删除特征集中的无关特征,再使用分类正确率或分类错误率度量标准选择最优特征子集。

表3 评价准则的简单比较

评价标准	分类性能	时间复杂度	特征	学习器
距离度量标准	一般	较低	连续型、离散型	分类、回归
一致性度量标准	一般	较低	离散型	分类
依赖性度量标准	一般	最低	连续型、离散型	分类、回归
信息度量标准	较好	较低	连续型、离散型	分类、回归
分类误差率或正确率	最好	最高	连续型、离散型	分类、回归

2.3 停止条件和结果验证

停止条件是判断特征选择过程是否结束的条件。停止条件一般与特征子集性能关系密切,设置阈值(指定的分类准确率、最大运行时间、最大迭代次数等)比较普遍,达到阈值便停止搜索,返回当前特征子集,另外,特征空间搜索完毕,特征选择过程自然就结束了。

结果验证是将最终返回的特征子集训练学习器,验证其有效性,保证原始特征集合可被其取而代之,简化后续分析。

3 特征选择分类

特征选择有很多种分类方式,根据有无类别特征,可以分为有监督、无监督特征选择算法;按照搜索策略,有全局最优、序列和随机搜索的特征选择算法;凭借评价准则包括距离度量、一致性度量、依赖性度量、信息度量以及分类正确率或错误率度量标准特征选择算法;依据特征选择和学习器的不同结合方式,囊括了过滤式(Filter)、封装式(Wrapper)、嵌入式(Embedded)和集成式(Ensemble)四种,这里主要介绍该分类方式。

3.1 过滤式

过滤式的特征选择算法和学习算法互不相干,特征选择是后者预处理过程,学习算法是前者验证过程。过滤式特征选择依照其特征选择框架的不同,又可以分为两类:基于特征排序和基于搜索策略。

3.1.1 基于特征排序

基于特征排序的Filter特征选择算法框架如图2所示。

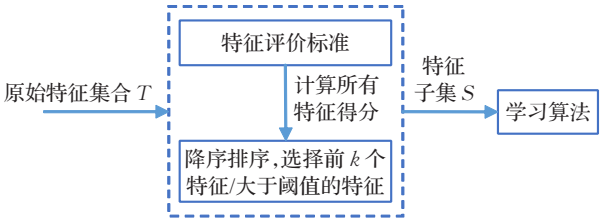

图2 基于特征排序的Filter特征选择框架

它采用具体的评价准则给每个特征打分,根据得分对特征降序排序,选择前 k 个特征作为特征子集(或者设置一个阈值,选择所有大于阈值的特征作为特征子集),最后用特征子集训练学习器验证子集的优劣。使用互信息度量特征重要性的BIF算法就是基于特征排序的Filter特征选择算法,这种方法简单快速,但只评估单个特征和类别的相关性,最终特征子集冗余度高。

基于特征排序的评价准则有Laplacian得分[50]、Constraint得分[51]、Fisher得分[52]、Pearson相关系数[53]、互信

息、MIC[49]等，换句话说，基于距离、一致性度量、依赖性和信息论的评价准则在基于特征排序的Filter算法中都可以使用。下面对一些评价准则作简要介绍：

(1)Laplacian得分

Laplacian得分所选特征方差大且同类(近邻)样本间变化小，即找到有区分度的特征。特征 f 的Laplacian得分定义为：

$$Laplacian(f)=\frac{\sum_{ij}(f_i-f_j)^2S_{ij}}{Var(f)} \tag{1}$$

式中，f_i 表示样本 i 在特征 f 上的取值；S_{ij} 表示权重矩阵 S 中的对应值，S 是对数据空间结构的模拟表达，描述了同类(近邻)样本两两之间的距离，两个同类(近邻)样本距离越大对应权重也越大，反之越小；$Var(f)$ 表示特征 f 的方差。由上式易知Laplacian得分越低，特征 f 越好。

Laplacian得分在分类、回归中皆适用。分类问题中，若特征 f 在同类样本中变化小，异类样本中变化大，则特征 f 较好，Laplacian得分较低；回归问题中，若特征 f 在近邻样本中变化小，其他样本中变化大，那么特征 f 较优，Laplacian得分较低。另外，无监督问题中亦可使用，使用方法同回归问题。

(2)Constraint得分

Constraint得分和Laplacian得分的思想大体一致，都是选择在同类样本变化小，异类样本变化大的特征，但Constraint得分不考虑方差，是有监督的特征选择算法，且只适合分类问题。Constraint得分首先定义must-link约束集 $M=\{(x_i,x_j)|x_i,x_j同类\}$ 和cannot-link约束集 $C=\{(x_i,x_j)|x_i,x_j异类\}$，然后使用约束集 M 和 C 对特征 f 评分，有两个不同的评分函数：

$$Constraint(f)^1=\frac{\sum_{(x_i,x_j)\in M}(f_i-f_j)^2}{\sum_{(x_i,x_j)\in C}(f_i-f_j)^2} \tag{2}$$

$$Constraint(f)^2=\sum_{(x_i,x_j)\in M}(f_i-f_j)^2-\lambda\sum_{(x_i,x_j)\in C}(f_i-f_j)^2 \tag{3}$$

式中，f_i 表示样本 i 在特征 f 上的取值，正则化系数 λ 平衡式(3)前后两项的贡献，$\lambda<1$。式(2)和式(3)均根据特征的约束保持能力打分，特征约束保持能力越好，得分越低，因为一个“好”特征可使must-link约束的两个样本彼此接近，使cannot-link约束的两个样本彼此远离。

(3)Fisher得分

Fisher得分根据特征对类别的可判定分离性打分，在分类问题中使用，与Laplacian得分、Constraint得分类似，“好”特征满足类内变化小，类间变化大，Fisher得分的评分函数为：

$$Fisher(f)=\frac{\sum_{i=1}^{c}n_i(\mu^i-\mu)^2}{\sum_{i=1}^{c}n_i(\sigma^i)^2} \tag{4}$$

式中，c 表示类别个数，n_i 表示第 $i(i=1,2,\cdots,c)$ 类样本的个数，μ^i 和 σ^i 表示第 $i(i=1,2,\cdots,c)$ 类样本中特征 f 的均值和方差，μ 表示特征 f 的均值。由上式可知，Fisher得分越高表示特征越好。

(4)Pearson相关系数

Pearson相关系数是计算两个特征线性相关程度的统计方法，特征 f_i、f_j 的Pearson相关系数为：

$$r(f_i,f_j)=\frac{Cov(f_i,f_j)}{\sqrt{Var(f_i)Var(f_j)}} \tag{5}$$

式中，$Cov(f_i,f_j)$ 表示特征 f_i、f_j 的协方差，$Var(f_i)$ 表示特征 f_i 的方差。通过计算特征与类别的相关系数对特征打分，得分越高，特征越好，在回归问题中亦可用，但是该准则只能计算线性相关，而更多时候特征与类别之间呈现复杂的非线性特性，所得特征子集的性能大打折扣。

(5)MIC

MIC具有很强的普适性，可识别任何函数关系，它打破了基于熵理论的评价准则只能处理离散型特征的瓶颈，因此MIC不仅可用于分类问题，还适用回归问题。MIC衡量特征和类别(或被解释特征)的相关性，值越大，相关性越高。特征 f_1、f_2 的MIC定义如下：

$$MIC(D)=\max_{XY<B(n)}M(D)_{X,Y}=\max_{XY<B(n)}\frac{I^*(D,X,Y)}{\ln(\min(X,Y))} \tag{6}$$

式中，$D=\{(f_{1i},f_{2i}),i=1,2,\cdots,n\}$ 是一个有序对集合，X 表示将 f_1 的值域划为 X 段，Y 表示将 f_2 的值域划为 Y 段，$XY<B(n)$ 表示网格数目不能大于 $B(n)$(数据总量的0.6或0.55次方)，分子 $I^*(D,X,Y)$ 表示不同 $X\times Y$ 网格划分下的互信息最大值(有多个)，分母 $\ln(\min(X,Y))$ 表示将不同划分下的最大互信息值归一化(还可以选择 $\lg(\cdot)$、$\mathrm{lb}(\cdot)$ 等对数函数)。

总而言之，使用该框架的特征选择算法效率高，因此在处理高维数据时，可在短时间内去除大量的无关特征。但是并非所有得分高的特征(强相关特征)组合在一起所获的特征子集的整体性能就一定好，其中有很多高度冗余特征，冗余特征对特征子集的整体性能有负面影响，另外，少量的弱相关特征是必要的。

3.1.2 基于搜索策略

基于搜索策略的Filter特征选择算法框架如图3所示。

它不单单使用简单的排序方式挑选子集，会运用一些启发式规则，有些结合前向搜索策略，每选择一个特

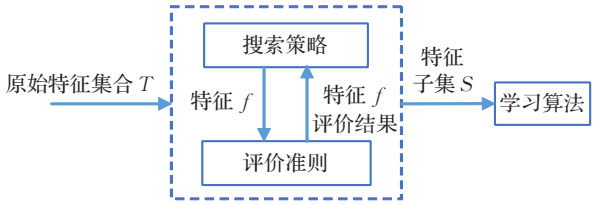

图3 基于搜索策略的Filter特征选择框架

征都对已选特征子集进行评价,但它有自己独特的准则衡量已选特征子集,如基于相关性的特征选择(Correlation based Feature Selection,CFS)算法用 $Merit_s$ 得分评价候选特征子集,若加入该特征使 $Merit_s$ 得分降低,那该特征便可剔除,基于搜索策略的Filter特征选择算法还有mRMR[18]、马尔科夫毯[54]等。该框架在特征选择过程中,对特征子集进行综合评价,一定程度上减少了特征子集的冗余度,并且由于不需要每增加一个特征就构建学习器进行特征子集评价,因此效率是比较高的。

(1)CFS

CFS估计特征子集而非单个特征的性能,它引入前向搜索策略,旨在选出强相关非冗余特征,CFS评价函数为:

$$Merit_s=\frac{kr_{fc}}{\sqrt{k+k(k-1)r_{ff}}} \tag{7}$$

式中,$Merit_s$ 表示特征子集 S 的类别区分能力,k 表示已选特征子集 S 中特征个数,r_{fc} 表示特征 f 与类别 c 的平均相关系数,r_{ff} 表示特征 f 与 S 中其他特征的平均冗余程度。$Merit_s$ 值越大表示特征子集 S 性能越好。

(2)mRMR

最小冗余最大相关算法使用增量搜索选择特征,与CFS算法相似,它可以最大化特征与类别的相关性,最小化特征间的冗余性,但它的两个评分函数如下:

$$J(f)^1=I(f;c)-\frac{1}{|S|}\sum_{s\in S}I(f;s) \tag{8}$$

$$J(f)^2=I(f;c)\Big/\frac{1}{|S|}\sum_{s\in S}I(f;s) \tag{9}$$

式中,S 表示已选特征子集,$|S|$ 表示特征个数,特征 f 和类别 c 的互信息 $I(f;c)$ 表示相关性,特征 f 和已选特征 s 的互信息 $I(f;s)$ 代表冗余性。虽然上式只衡量单个特征得分,但mRMR还使用代价函数评价特征子集,因此子集性能较好。

(3)马尔科夫毯

马尔科夫毯主要用于去除冗余特征,已知全集 F,若给定特征子集 $M(M\in F)$,特征 f 独立于 $F-M-\{f\}$ 和类别 c,则 M 是 f 的马尔科夫毯,表示为:$f\perp F-M-\{f\}|M$。因此,当特征子集 M 存在时,特征 f 对分类无贡献,是冗余特征。

由于马尔科夫毯计算冗余特征复杂度太高,因此实际应用中皆是运用近似马尔科夫毯计算,特征 f_i 是 f_j 的近似马尔科夫毯的条件是:

$$J(f_i,c)>J(f_j,c)\ \text{且}\ J(f_i,f_j)>J(f_j,c) \tag{10}$$

式中,$J(f_i,c)$ 表示特征 f_i 和类别 c 的相关性,$J(f_i,f_j)$ 表示 f_i、f_j 的相关性,函数 $J(\cdot)$ 可选,条件熵、互信息、对称不确定性 SU、MIC等先后被提出。上式中的 f_j 视为冗余特征。

3.1.3 基于特征排序+搜索策略

基于特征排序+搜索策略的Filter特征选择方法通常包含两大步骤:第一步使用基于特征排序的Filter算法去除无关特征,第二步使用基于搜索策略的Filter算法删除冗余特征,如图4所示,这类框架通常用于处理高维数据集,可较迅速地获取高性能的特征子集,这类算法有快速过滤式特征选择算法FCBF、FCBF-MIC[54],文献[55]中提出的基于特征聚类和联合对称不确定性的算法JSU-FCBF,文献[56]中提出的利用近似马尔科夫毯的最大相关最小冗余特征选择算法nmRMR等。

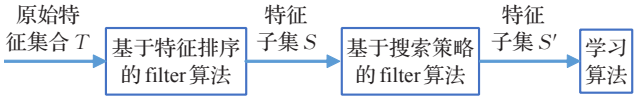

图4 基于特征排序+搜索策略的Filter特征选择框架

3.2 封装式

Wrapper方法是一种特征选择过程与学习算法结合的特征选择方法,例如张戈等[57]提出的AB-CRO算法就是基于Wrapper框架的。Wrapper将选用的学习器封装成黑盒,根据它在特征子集上的预测精度评价所选特征的优良,并采用搜索策略调整子集,最终获得近似的最优子集,如图5所示。

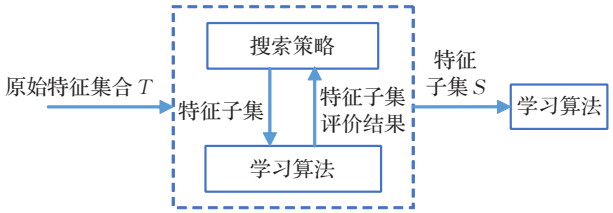

图5 Wrapper特征选择框架

封装式特征选择方法每每由两部分组成,即搜索策略和学习算法,搜索策略前面已经介绍了,不再复述,学习算法主要用来评判特征子集的优劣,学习算法的选取不受限制,分类问题可使用支持向量机、k 近邻等,如果数据维度高、高维小样本时可选用支持向量机,k 近邻适合样本量不大、维度不太高的情况;回归问题可选择最小二乘回归、偏最小二乘回归(PLS)、Lasso等,如果数据样本量多于特征数可使用最小二乘回归,PLS可实现多自变量对多因变量、高维小样本数据的回归建模,Lasso亦可用于高维小样本数据且其本身就有特征选择的功能。

该框架使用特定的学习算法得到的特征子集效果非常好。但是特征子集的性能受特定的学习算法影响，容易“过拟合”，如使用支持向量机选取的特征子集在 k 近邻上的效果与之相差甚远；使用不同的学习算法，得到的特征子集也不一样，所以特征子集的稳定性和适应性较差；另外，由于每增加一个特征就要构造学习器对特征子集进行评价，因此该框架时间复杂度高，不适合高维数据集。

3.3　嵌入式

嵌入式特征选择算法嵌入在学习算法当中，当分类算法训练过程结束就可以得到特征子集。嵌入式特征选择方法可解决基于特征排序的Filter算法结果冗余度过高的问题，还可以解决Wrapper算法时间复杂度过高的问题，它是Filter和Wrapper的折中。

嵌入式特征选择算法没有统一的流程框架图，不同的算法框架各异。分类决策树是经典的嵌入式特征选择算法，特征选择框架如图6所示，包括有：ID3[58]、C4.5[59]、CART[60]等算法，训练用到的特征便是特征选择的结果，可运用到分类、回归问题中。

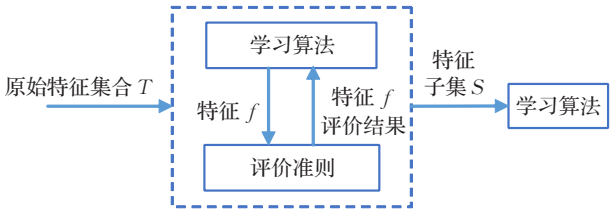

图6　分类决策树的特征选择框架

另一类基于L1正则项，有基于L1正则项的最小二乘回归方法Lasso[61]，均分式Lasso[62]，K-split Lasso[63]，采用迭代思想的GSIL[64]和迭代式Lasso[65]，与序贯思想相结合的SLasso[66]，融合L2正则项的弹性网Lasso[67]，采用有监督组的SGLasso[68]，Zhu等针对高维数据提出的基于L1正则项的SVM算法[69]等。以上算法都可用于回归问题的特征选择，L1正则项的性质会使回归系数朝着0收缩，并且较小的系数可能会压缩为0，因此那些不为0的系数所对应的特征就是最终的特征子集。

3.4　集成式

集成式特征选择算法借鉴集成学习思想，它训练多个特征选择方法，并整合所有特征选择方法的结果，可获得比单个特征选择方法更好的性能。正巧决策树可以特征选择，所以决策树作为基分类器的集成学习方法随机森林(RF)就是一种集成式特征选择算法。通过引入Bagging思想，很多特征选择算法可改进为集成式。例如文献[64]提出的特征选择算法ECGS-RG、文献[70]提出的四阶段特征选择算法本质上皆是集成算法，可提高算法的稳定性且适合高维小样本数据。

3.5　几类特征选择算法的优缺点

Filter算法效率高，尤其是基于排序的算法，它适合各种数据类型，因为基于距离、一致性度量、依赖性和信息论的评价准则都可套用该框架。若是分类问题，可以选择Laplacian得分、Constraint得分、Fisher得分、MIC等指标；Pearson相关系数、MIC以及依赖性准则的指标可用于回归问题；若是离散型特征，建议选择互信息、对称不确定性、MIC等信息论准则；若是连续性特征，MIC、Pearson相关系数和一些依赖性评价指标可供选择。基于排序的算法缺点也很明显，它只估计单个特征的得分，未评估特征子集性能，得到的子集冗余度高。

基于搜索策略的Filter算法可得到性能较好的子集，它运用启发式规则搜索，并根据子集性能挑选特征，能去除一些冗余特征，但其时间复杂度比基于排序算法更高。如果数据维度不算太高，可使用CFS、mRMR等可选出性能较好且一定程度去冗余的算法，CFS多用于分类数据，对特征类型要求不高，离散型、连续型特征皆可，mRMR因其使用互信息计算，一般只用于分类的离散型数据；若是成千上万的高维数据，近似马尔科夫毯将会是一个好选择，选择不同的函数构建近似马尔科夫毯可使其灵活运用于分类、回归问题和离散、连续型特征中，它是去除冗余特征的利器，但会有过度去冗余的风险，也即很可能删掉有用的信息特征以致特征子集性能下降，但这在高维数据中较常见，因为从成千上万维降至几十维丢失一些信息是可以理解的。

基于特征排序+搜索策略的Filter特征选择方法先去除无关特征，再去除冗余特征，因为生成子集过程不涉及构建分类器，因此时间复杂度不算太高，是处理高维、高维小样本数据的可选框架。

Wrapper算法选择的子集性能较优，它是搜索策略和学习器结合使用的，性能好坏和搜索策略、学习器息息相关。若是选用序列搜索方式，则时间复杂度高，易发生过拟合，不适合高维数据且只能获取局部最优解；随机搜索方式可以用于高维数据并且获得近似最优解，但降维效果会差一些。学习器将决定特征子集的性能，面对高维数据，SVM比 k 近邻更加适合。

Embedded算法数量较少，这取决于学习器的特性，某些学习器天然具有特征选择的功能。Embedded算法效率高，特征子集性能优异但只针对其本身，且易出现过拟合。一般分类问题中可选用决策树，而正因为L1正则项可将回归系数压缩为0的性质，吸引了众多学者，使其在回归问题的特征选择研究中大放异彩。

Ensemble算法针对特征选择方法的稳定性进行研究，特征选择算法大都对数据分布敏感，若是特征选择的训练集发生改变，即便同一算法，结果也大相径庭，这是一个非常严重的问题，特征子集无法复现，近年来学者们提出的许多集成式算法略有成效，但都有较高的时间要求。

各类特征选择算法的优缺点见表4，总体来说，Filter

表4 各类特征选择算法的优缺点

类别		优点	缺点
Filter	基于特征排序	效率高	特征子集的冗余度高 依赖具体的度量标准
	基于搜索策略	效率较高 特征子集冗余度较低	依赖具体的度量标准 不适合高维数据集
	特征排序+搜索策略	效率较高 特征子集冗余度较低 适合高维数据集	依赖具体的度量标准
Wrapper	基于序列搜索	效率较高 特征子集性能较好	只能获取局部最优解 依赖具体的学习算法 容易过拟合 不适合高维数据集
	基于随机搜索	可获取近似最优解 特征子集性能好	效率较低 不适合高维数据集
Embedded		效率较高 特征子集性能较好 可处理高维数据集	依赖具体的学习算法 可能出现过拟合
Ensemble		特征子集稳定性好	效率较低

时间复杂度低,但是特征子集的选取依赖具体的度量标准;Wrapper的特征子集性能较好,但是特征子集的性能对学习算法依赖性高,易“过拟合”,且不适合高维数据集;Embedded效率较高,特征子集性能较好,也可以处理高维数据集,但是特征子集的选择依赖具体的学习算法且可能出现“过拟合”的问题;Ensemble可解决以上三种算法特征子集不稳定的问题,但通常时间复杂度较高。

4 结束语

本文首先阐述了特征选择框架中的四个过程,然后按照特征选择和学习算法的关联方式对特征选择进行分类。特征选择算法数量庞大,种类繁多,但缺陷依然无可避免,研究对象的日益复杂使现有算法性能不佳,因此,如何针对实际问题设计方案亟待解决。

(1)设计高维小样本数据的特征选择方法

高维小样本数据集的处理方式大都使用多阶段特征选择算法,第一个阶段去除无关特征,第二阶段删减冗余特征,如FCBF算法。但是它经过特征选择之后预测精度可能下降,如何在特征子集预测精度不下降的前提下设计高维小样本数据的特征选择方法具有一定的挑战性。

(2)特征选择算法的稳定性

现有的特征选择算法依赖具体的度量标准或者特定的学习算法,因此使用不同的学习算法进行数据分析通常需要进行多次特征选择;另外,特征选择算法对数据分布敏感,若训练集发生变化,即便同一算法,特征子集也千差万别。因此,如何提高特征选择算法的稳定性至关重要。

(3)回归分析中的特征选择算法

回归问题中因其连续型自变量和因变量导致很多算法不适用,因此回归问题中的特征选择方法值得研究。

(4)冗余特征的去除

现有的特征选择算法大都可以删除无关特征并且在一定程度上去冗余,或者冗余特征过度删除导致丢失大量信息,因此如何准确有效地去除冗余特征非常重要。

参考文献:

[1] Kozodoi N,Lessmann S,Papakonstantinou K,et al.A multi-objective approach for profit-driven feature selection in credit scoring[J].Decision Support Systems,2019,120:106-117.

[2] Labbé M,Martínez-Merino L I,Rodríguez-Chía A M.Mixed integer linear programming for feature selection in support vector machine[J].Discrete Applied Mathematics,2019,261:276-304.

[3] Jayaprakash A,KeziSelvaVijila C.Feature selection using Ant Colony Optimization(ACO) and Road Sign Detection and Recognition(RSDR) system[J].Cognitive Systems Research,2019,58:123-133.

[4] Blum A L,Langley P.Selection of relevant features and examples in machine learning[J].Artificial Intelligence,1997,97(1/2):245-271.

[5] Francois D.High-dimensional data analysis:optimal metrics and feature selection[D].Louvain-la-Neuve:University Catholique de Louvain,2006.

[6] Jolliffe I T.Pincipal component analysis[J].Journal of Marketing Research,2002,25(4):513.

[7] Liu C.Gabor-based kernel PCA with fractional power polynomial models for face recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2004,26(5):572-581.

[8] Fisher R A.The use of multiple measurements in taxonomic problems[J].Annals of Eugenics,1936,7:179-188.
[9] Baudat G,Anouar F.Generalized discriminant analysis using a kernel approach[J].Neural Computation,2000,12:2385-2404.
[10] Hyvarinen A,Oja E,Karhunen J.Independent component analysis[M].New York:Wiley,2001.
[11] Bach F R,Jordan M I.Kernel independent component analysis[J].Journal of Machine Learning Research,2002,3:1-48.
[12] Cox T,Cox M.Multidimensional scaling[M].London:Chapman & Hall,1994.
[13] Kira K,Rendell L A.The feature selection problem:traditional methods and a new algorithm[C]//Proceedings of the 10th National Conference on Artificial Intelligence,San Jose,CA,July 12-16,1992.[S.l.]:AAAI Press,1992:129-134.
[14] John G H,Kohavi R,Pfleger K.Irrelevant features and the subset selection problem[C]//Proceedings of the 11th International Conference on Machine Learning,1994:121-129.
[15] Koller D,Sahami M.Toward optimal feature selection[C]//Proceedings of International Conference on Machine Learning,Bari,1996:284-292.
[16] Dash M,Liu Huan.Feature selection for classification[J].Intelligent Data Analysis,1997,1(3):131-156.
[17] Saeys Y,Inza I,Larranaga P.A review of feature selection techniques in bioinformatics[J].Bioinformatics,2007,23(19):2507-2517.
[18] Peng H,Long F,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.
[19] 赖学方,贺兴时.最小冗余最大分离准则特征选择方法[J].计算机工程与应用,2017,53(12):70-75.
[20] Yu L,Liu H.Efficient feature selection via analysis of relevance and redundancy[J].Journal of Machine Learning Research,2004,5:1205-1224.
[21] Sun Z H,Bebis G,Miller R.Object detection using feature subset selection[J].Pattern Recognition,2004,37(11):2165-2176.
[22] Somol P,Pudil P,Kittler J.Fast branch & bound algorithms for optimal feature selection[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2004,26(7):900-912.
[23] Davies S,Russell S.NP-completeness of searches for smallest possible feature sets[C]//Proceedings of the 1994 AAAI Fall Symposium on Relevance,1994:37-39.
[24] 张永波,游录金,陈杰新.基于模拟退火的多标记数据特征选择[J].计算机工程与设计,2011,32(7):2494-2496.
[25] 叶志伟,郑肇葆,万幼川,等.基于蚁群优化的特征选择新方法[J].武汉大学学报(信息科学版),2007(12):1127-1130.
[26] Wutzl B,Leibnitz K,Rattay F,et al.Genetic algorithms for feature selection when classifying severe chronic disorders of consciousness[J].PloS one,2019,14(7).
[27] 周丹,吴春明.基于改进量子进化算法的特征选择[J].计算机工程与应用,2018,54(1):146-152.
[28] 张翠军,陈贝贝,周冲,等.基于多目标骨架粒子群优化的特征选择算法[J/OL].计算机应用,2019:1-7[2019-08-20].http://kns.cnki.net/kcms/detail/51.1307.tp.20180719.1001.002.html.
[29] Fodor I K.A survey of dimension reduction techniques Number[R].Lawrence Livermore National Laboratory,US Department of Energy,2002.
[30] Kira K,Rendell L A.A practical approach to feature selection[C]//Proceedings of the Ninth International Workshop on Machine Learning(ML 1992),Aberdeen,Scotland,UK,July 1-3,1992.[S.l.]:Morgan Kaufmann Publishers Inc,1992:249-256.
[31] Kononenko I.Estimating attributes:analysis and extensions of RELIEF[J].Machine Learning,1994,784:171-182.
[32] Arauzo-Azofra A,Benitez J M,Castro J L.Consistency measures for feature selection[J].Journal of Intelligent Information System,2008,30:273-292.
[33] Dash M,Liu H.Consistency-based search in feature selection[J].Artificial Intelligence,2003,151(1/2):155-176.
[34] Almuallim H,Dietterich T G.Learning with many irrelevant features[C]//Proceedings of 9th National Conference on Artificial Intelligence,Menlo Park,1992:547-552.
[35] Liu H,Setiono R.A probabilistic approach to feature selection-a filter solution[C]//Proceedings of International Conference on Machine Learning,Bari,1996:319-327.
[36] Devijver P A,Kittler J.Pattern recognition:a statistical approach[M].London:Prentice Hall,1992.
[37] Hall M A.Correlation-based feature subset selection for machine learning[D].Hamilton,New Zealand:University of Waikato,1999.
[38] Zhang D,Chen S,Zhou Z H.Constraint score:a new filter method for feature selection with pairwise constraints[J].Pattern Recognition,2008,41(5):1440-1451.
[39] Jain A K,Duin R P W,Mao J.Statistical pattern recognition:a review[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(1):4-37.
[40] Battiti R.Using mutual information for selecting features in supervised neural net learning[J].IEEE Transactions on Neural Networks,1994,5(4):537-550.
[41] Kwak N,Choi C H.Input feature selection by mutual information based on Parzen window[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24(12):1667-1671.
[42] Huang J,Cai Y,Xu X.A hybrid genetic algorithm for

feature selection wrapper based on mutual information[J]. Pattern Recognition Letters,2007,28:1825-1844.

[43] Novovicova J,Somol P,Haindl M,et al.Conditional mutual information based feature selection for classification task[C]// Proceedings of the 12th Iberoamericann Congress on Pattern Recognition,Valparaiso,Chile,November 13-16 2007:417-426.

[44] Qu G,Hariri S,Yousif M.A new dependency and correlation analysis for features[J].IEEE Transactions on Knowledge and Data Engineering,2005,17(9):1199-1207.

[45] Estevez P A,Tesmer M,Perez C A,et al.Normalized mutual information feature selection[J].IEEE Transactions on Neural Networks,2009,20(2):189-201.

[46] 刘华文.基于信息熵的特征选择算法研究[D].长春:吉林大学,2010.

[47] Fleuret F.Fast binary feature selection with conditional mutual information[J].Journal of Machine Learning Research,2004,5:1531-1555.

[48] Meyer P E,Schretter C,Bontempi G.Information-theoretic feature selection in microarray data using variable complementarity[J].IEEE Journal of Selected Topics in Signal Processing,2008,2(3):261-274.

[49] Reshef D N,Reshef Y A,Finucane H K,et al.Detecting novel associations in large data sets[J].Science,2011,334(6062).

[50] He X,Cai D,Niyogi P.Laplacian score for feature selection[C]//Advances in Neural Information Processing Systems,2006:507-513.

[51] Zhang D,Chen S,Zhou Z H.Constraint score:a new filter method for feature selection with pairwise constraints[J].Pattern Recognition,2008,41(5):1440-1451.

[52] Bishop C M.Neural networks for pattern recognition[M]. Oxford:Oxford University Press,1995.

[53] van't Veer L J,Dai H,van de Vijver M J,et al.Gene expression profiling predicts clinical outcome of breast cancer[J].Nature,2002,415(6871):530-536.

[54] 孙广路,宋智超,刘金来,等.基于最大信息系数和近似马尔科夫毯的特征选择方法[J].自动化学报,2017(5).

[55] 郝志强.基于联合对称不确定性的特征选择算法研究[D].辽宁大连:大连理工大学,2017.

[56] 张俐,王枞,郭文明.利用近似马尔科夫毯的最大相关最小冗余特征选择算法[J].西安交通大学学报,2018,52(10):147-151.

[57] 张戈,王建林.基于混合ABC和CRO的高维特征选择方法[J/OL].计算机工程与应用,2019:1-13[2019-09-03].http://kns.cnki.net/kcms/detail/11.2127.tp.20190322.1838.012.html.

[58] Quinlan J R,Learning efficient classification procedures and their application to chess end games[C]//Michalski R S,Carbonell J G,Mitchell T M.Machine Learning:An Artificial Intelligence Approach.Los Altos:Morgan Kaufmann,1983:463-482.

[59] Quinlan J R.C4.5:programs for machine learning[M].[S.l.]:Morgan Kaufmann Publishers Inc,1992.

[60] Everitt B S.Classification and regression trees[M]//Encyclopedia of statistics in behavioral science.[S.l.]:John Wiley & Sons,Ltd,2005.

[61] Tibshirani R.Regression shrinkage and selection via the Lasso[J].Journal of the Royal Statistical Society Series B-Methodological,1996,58(1):267-288.

[62] 施万锋,胡学钢,俞奎.一种面向高维数据的均分式Lasso特征选择方法[J].计算机工程与应用,2012,48(1):157-161.

[63] 张靖,胡学钢,张玉红,等.K-split Lasso:有效的肿瘤特征基因选择方法[J].计算机科学与探索,2012,6(12):1136-1143.

[64] 张靖.面向高维小样本数据的分类特征选择算法研究[D].合肥:合肥工业大学,2014:35-52.

[65] 施万锋,胡学钢,俞奎.一种面向高维数据的迭代式Lasso特征选择方法[J].计算机应用研究,2011,28(12):4463-4466.

[66] Luo S,Chen Z.Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space[J]. Journal of the American Statistical Association,2014,109(507):1229-1240.

[67] Zou H,Hastie T.Regularization and variable selection via the elastic net[J].Journal of the Royal Statistical Society,2005,67(2):301-320.

[68] Ma S,Song X,Huang J.Supervised group Lasso with applications to microarray data analysis[J].BMC Bioinformatics,2007,8(1):1-17.

[69] Zhu J,Rosset S,Hastie T,et al.1-norm support vector machines[C]//Proceedings of the 16th International Conference on Neural Information Processing Systems,2003.

[70] Pehlivanlı A C.A novel feature selection scheme for high-dimensional data sets:four-staged feature selection[J].Journal of Applied Statistics,2015,43(6):1-15.