Journal of Computer Applications ISSN 1001-9081 2017-09-10
计算机应用,2017,37(9):2433 -2438,2448 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2017)09-2433-06 DOI:10.11772/j.issn.1001-9081.2017.09.2433

高维小样本分类问题中特征选择研究综述

王 翔[1,2*],胡学钢[1]

(1.合肥工业大学 计算机信息学院,合肥 230009; 2.安徽省科学技术情报研究所 文献情报分析中心,合肥 230011)

(* 通信作者电子邮箱 wangxiang@ahinfo.gov.cn)

摘 要:随着生物信息学、基因表达谱微阵列、图像识别等技术的发展,高维小样本分类问题成为数据挖掘(包括机器学习、模式识别)中的一项挑战性任务,容易引发"维数灾难"和过拟合问题。针对这个问题,特征选择可以有效避免维数灾难,提升分类模型泛化能力,成为研究的热点,有必要对国内外高维小样本特征选择主要研究情况进行综述。首先分析了高维小样本特征选择问题的本质;其次,根据其算法的本质区别,重点对高维小样本数据的特征选择方法进行分类剖析和比较;最后对高维小样本特征选择研究面临的挑战以及研究方向作了展望。

关键词:特征选择;高维数据;小样本学习;信息过滤;支持向量机

中图分类号:TP391.4 **文献标志码**:A

Overview on feature selection in high-dimensional and small-sample-size classification

WANG Xiang[1,2*], HU Xuegang[1]

(1. *School of Computer and Information, Hefei University of Technology, Hefei Anhui* 230009, *China*;
2. *Literature Information Analysis Department, Anhui Institute of Scientific and Technical Information, Hefei Anhui* 230011, *China*)

Abstract: With the development of bioinformatics, gene expression microarray and image recognition, classification on high-dimensional and small-sample-size data has become a challenging task in data ming, machine learning and pattern recognition as well. High-dimensional and small-sample-size data may cause the problem of "curse of dimensionality" and overfitting. Feature selection can prevent the "curse of dimensionality" effectively and promote the generalization ability of classification mode, and thus become a hot research topic. Accordingly, some recent development of world-wide research on feature selection in high-dimensional and small-sample-size classification was briefly reviewed. Firstly, the nature of high-dimensional and small-sample feature selection was analyzed. Secondly, according to their essential difference, feature selection algorithms for high-dimensional and small-sample-size classification were divided into four categories and compared to summarize their advantages and disadvantages. Finally, challenges and prospects for future trends of feature selection in high-dimensional small-sample-size data were proposed.

Key words: feature selection; high-dimensional data; small-sample-size learning; information filtering; Support Vector Machine (SVM)

0 引言

随着科学研究的不断进步,数据挖掘领域需处理的对象越来越复杂,其数据维度也在急剧增加,在图像识别、文本分类、生物信息学、基因微阵列分析等大规模数据挖掘应用中出现了一个被称为高维小样本学习的研究热点。

高维小样本数据,是指具备数据维数特别高,样本绝对数量很少或样本数远小于数据维数特征的数据(也有文献称为"Large p small n"数据[1],p 代表特征数,n 代表样本数)。

高维小样本数据上的分类问题是机器学习中的难点,针对不同的应用领域衍生出了高维小样本、高维稀疏、高维不平衡等多个研究热点。较高的维数是获得问题准确描述的有力保障,而难以避免地会含有大量冗余、不相关和噪声特征,由于高维小样本数据的这些特点,容易引发"维数灾难"(Curse of Dimensionality),即随着维数增高,计算复杂度显著增高而分类器的性能急剧下降;又因为其样本数量很少,传统的分类学习方法效能严重下降,容易出现过拟合(Over Fitting),无法进行有效的分类或识别;近年来,高维小样本导致的特征选择模型的不稳定性也引起了重视。为了消除或减轻维数灾难,同时提升分类器的泛化能力,降维成为重要途径。

在不同应用领域的高维小样本降维研究中,特征抽取与特征选择是两类非常重要的技术。特征抽取主要是是将高维数据映射到特定的低维空间,而特征选择可以看作从初始特征空间搜索出一个最优特征子集的过程。

从算法原理上分析,特征抽取是一种基于变换的方法,原数据中的不相关和冗余的特征均在降维中产生作用,影响了分类性能,且新的低维特征空间中特征失去原有的物理解释,对某些高维小样本数据分类问题(如癌症基因分析)而言,很难接受。特征选择并不改变原特征空间,只是选择一些分辨力好的特征,组成一个新的低维空间,可以保留原始特征空间大部分性质,对于高维小样本而言,特征选择可以去除不相关特征和冗余特征,在一定程度上将噪声数据对分类器性能的

收稿日期:2017-03-27;修回日期:2017-04-21。 基金项目:国家973计划项目(2016YFC0801406);国家自然科学基金资助项目(61673152);安徽省自然科学基金资助项目(1408085QF136)。

作者简介:王翔(1982—),男,安徽合肥人,博士研究生,主要研究方向:数据挖掘、人工智能、情报分析; 胡学钢(1962—),男,安徽合肥人,教授,博士,主要研究方向:数据挖掘、人工智能、大数据分析。

影响降到最低,且选择的特征可解释性较好。

特征选择已成为高维小样本数据分类问题中的关键性步骤,一直是机器学习和数据挖掘研究的热点之一,新的算法也不断地被提出,本文对当前已进行的研究进行综述,尝试从原理上分析这些算法的区别与联系,总结各自的优点与不足,并对未来高维小样本特征选择的研究进行展望。

1 高维小样本中的特征选择

为了便于比较和分析,本文按照评价函数的不同将高维小样本特征选择方法分为 Filter(筛选法)、Wrapper(封装法)、Embedded(嵌入式)以及 Ensemble(集成法)四类,在实际高维小样本分类应用中,Embedded 方法因其适合处理小样本问题而受到了诸多研究人员的关注,本文也将针对该类方法作重点分析。

1.1 Filter(筛选法)

Filter(筛选法)通过分析特征子集内部的特点来衡量特征的分类能力,与后面的采用何种分类器无关,这类方法通常需要评价特征相关性的评分函数和阈值判别法来选择出得分最高的特征子集。通过文献调研,根据选择特征子集方式的不同,可以继续划分为基于特征排序(Feature Ranking)和基于特征空间搜索(Space Search)两类。

基于特征排序的方法,其主要思想[2]是:

1)使用评分函数(Scoring Function)对每个特征进行评分,并将所有特征按照得分的降序排列;

2)对每个特征得分进行显著性检验(如 p-value 等);

3)通过预先设置的阈值选择排序前列的具有显著统计学意义的特征;

4)验证选择的最优特征子集,通常使用 ROC(Receiver Operating Characteristic)曲线、分类正确率、组相关系数、稳定性等。

基于特征排序方法的核心就是评分函数,表1列举了高维小样本分类应用中出现的基于度量样本群分布之间的差异、基于信息论、基于相关性标准等三类热门评分函数。

表1 按评分函数分类的基于特征排序方法

Tab. 1 Category of feature ranking by scoring function

名称	算法描述	应用
基于度量样本群分布之间差异的评分函数	通过统计方法度量同类样本间几何关系度量值与异类样本间几何关系度量值的分布差异评价特征	t-test[3] Fold-Change Ratio and Difference[3-4] Z-score[5] 贝叶斯框架[6-8] 概率密度[9-10]
基于信息论的评分函数	利用信息熵等指标度量目标特征所包含的信息量来评分	互信息[11-12] 信息增益[13] 信噪比[14-15]
基于相关性标准的评分函数	通过测量目标特征与类别之间的相关性来衡量目标特征的重要程度,与类标签的相关度越高,目标特征分辨力越好	泊松系数[16] Kendall 排序[17] Fisher 分析[18]

基于特征空间搜索法主要是采用一种优化策略从整个特征集合中选出包含最多信息并且达到最小冗余的特征子集。在特定领域,如致病基因的准确发现有一些研究,如基于关联规则(Correlation-based Feature Selection, CFS)、最大相关最小冗余(Maximum Relevance Minimum Redundancy, MRMR)等,表2给出了上述基于特征空间搜索法的几类主流方法。

表2 基于特征空间搜索方法

Tab. 2 Space search-based feature selection

名称	算法描述
CFS[19]	采用“heuristic merit”选择出满足集合内所有特征都与类标签最大相关且相互间相关度最小要求的特征子集
MRMR[20]	采用互信息或 F 值(p-value)指标来找到满足具有最小冗余性和最大相关性的特征集合
马尔可夫毯[21]	通过排除与类标签相互独立的特征来构建特征子集
ReliefF[22]	随机搜索某个样本,通过比较其同类标签中的 K 个最近邻与异类标签中的 K 个最近邻来度量与类标签最相关的特征

基于特征排序的方法多为单变量方法,每次考虑单个特征的影响,选择与类标签最相关的特征,对高维小样本来说具有较低的计算复杂度,但在某些应用领域如基因微阵列数据中,因忽略了特征间的相互关系,直接应用分类精度较为一般;而基于特征空间搜索为多变量的方法,这类算法不但需要考虑特征子集与类标签的相关性,还需要考虑特征子集之间的相关性,通常分类正确率较高,但在高维条件下寻找最优子集过程的计算复杂度较高。

1.2 Wrapper(封装法)

Wrapper 方法是一种与分类模型结合的特征选择方法,使用某个分类模型封装成黑盒,根据这个分类器在特征子集上的结果好坏来评价所选择的特征,并采取某些优化的搜索策略对子集进行调整,最终获得近似的最优子集。N 个特征的数据集,可能的特征子集数为 2^N 个,发现最优特征子集已经被证明是 NP-Hard,因此高维小样本特征选择中 Wrapper 研究热点并不局限于采用何种分类模型作为评价准则(通常使用遗传算法(Genetic Algorithm, GA)[23]、支持向量机(Support Vector Machine, SVM)[24]、K 最近邻(K-Nearest Neighbor, KNN)[25]构建分类模型),特征子集搜索策略成为研究热点。

根据文献[26],Wrapper 方法可以被粗略划分为顺序搜索与启发式搜索两类(有文献将其分类为确定性与随机性[27]):

1)顺序搜索算法。

顺序搜索算法从一个空的特征子集开始,通过不断增加(或删除)特征直到特征子集能使评价函数得到最好表现,现实中会引入一些停止搜索的标准加速特征子集的选择,确保评价函数持续增加并达到最好表现时,所选择特征子集具有最少数目。图1给出了常规顺序搜索算法的发展脉络。

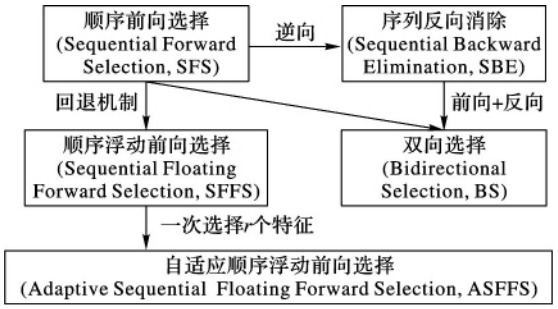

图1 顺序搜索算法发展脉络

Fig. 1 Development of sequential search algorithms

顺序搜索策略不断增加(或删除)单个特征,避免了完全搜索计算复杂度过高的问题,但选择出的特征子集很难避免

所选特征之间的高度相关性[27]；自适应的顺序浮动选择方法(Adaptive Sequential Floating Forward Selection，ASFFS)[28]引入了参数 r 用于指定加入特征的数量，并使用参数 o 去除属性；同时，为了提升时间性能，Nakariyakul 等[29]提出一种新的特征选择方法——“增 L 去 R”，每次从候选特征集中添加(或删除) L 个特征，考虑入选(或删除)的 L 个特征之间相关性，然后删除(或添加)与类标签最不相关(或最相关)的 R 个特征，有效提升了算法的时间性能。

2) 启发式搜索算法。

启发式搜索算法所开始搜索的初始特征子集是从完整的特征候选集中随机生成，通过启发式规则逐步接近最优解，这种方法在搜索时具有很强的不确定性，但随着算法的运行，使用这种策略获得的特征子集的质量也能满足需要。在高维小样本特征选择中常用的启发式搜索方法有 GA 和粒子群优化(Particle Swarm Optimization，PSO)算法[30]等，通过设定迭代次数的阈值，从而降低需要搜索的特征子集的数量，通过调整 GA 的参数和遗传算子，还可以获得更多的应用。Cordón 等[31]使用进化遗传算法的方法对图像识别领域中的特征选择问题进行了尝试。

因为顺序搜索算法的特点，无法从已抛弃的特征中进行二次选择，也无法抛弃已选择的特征，容易陷入被称为嵌套影响(Nesting Effect)的局部最优情况，启发式搜索算法可以有效解决这些问题，并且有文献表明，采用并行设计启发式搜索算法的计算复杂度比顺序选择算法显著降低[32]。

1.3　Embedded(嵌入式)

Embedded(嵌入式)方法的出现主要是为了解决 Wrapper(封装法)在处理不同数据集时，分类模型需要重构代价高等问题。如果严格区分，它与 Wrapper 的不同在于，Embedded 将特征选择与分类模型的学习过程结合，即在分类器的训练过程中包含了特征选择功能，由于其高效的时空性能及较好的分类精度，逐渐成为高维小样本特征选择的热点方向，其中，有两类方法成为非常热门的研究对象：一是以 SVM 为基础的模型；二是以套索算法 Lasso(Least Absolute absolute Shrinkage shrinkage and Selection selection Operator)为代表的正则化组稀疏模型。

由于 SVM 可以根据有限的样本在模型的复杂性和学习能力之间寻求最佳平衡，弱化了对数据正态分布的要求，同时对维数灾难不敏感，可以有效剔除冗余特征，具有较好的泛化性能，因而被广泛用于处理高维小样本特征选择。

Guyon 等[33]采用基于递归特征的后项搜索剔除思想(Recursive Feature Elimination，RFE)，提出 SVM-RFE 方法，该方法将 SVM 超平面的每个维度对应高维小样本数据集里的每个特征，从而每个维度权重的绝对值可以用来度量对应数据特征的重要性，即通过权重对特征进行降序排序。从排序后的特征集合开始，每次删除排名靠后的一个特征，迭代直到该特征集合为空，一般来说最先被删除的多为噪声或冗余特征，最后被删除的特征一般具有较强的区分能力，由于该方法在处理高维小样本数据方面的优势及贪婪算法带来的计算复杂度，围绕该方法研究人员提出了许多改良方法，其中有按比例删除特征的[34-35]，采用前向序列思想的[36-37]，采用模糊聚类的[38]，基于 Relief 的[39]以及采用了粒子群算法的[40]，图 2 给出了部分基于 SVM-RFE 的算法分类。

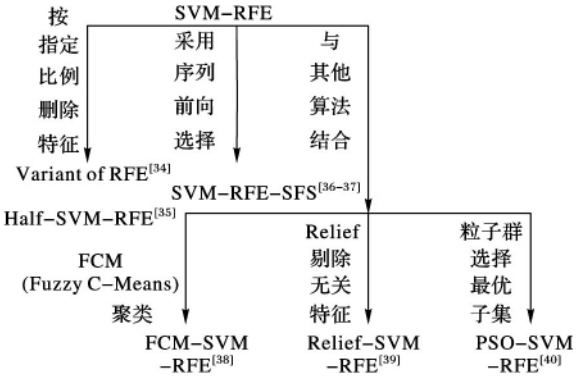

图 2　基于 SVM-RFE 的算法分类

Fig. 2　Categories of SVM-RFE based algorithms

除基于 SVM 的方法之外，正则化稀疏模型可以将大量冗余或噪声特征去除，同时保留与类标变量最相关的特征，选择的特征子集具有很好的可解释性，也获得了许多研究人员的关注。其代表算法有 1996 年 Tibshirani[41]提出的 Lasso，其基本思想是在最小二乘估计中加入系数的绝对值作为惩罚项，使系数的绝对值之和小于等于某一个阈值来最小化残差平方，能产生趋于 0 的回归系数，使得与目标关系弱的预测变量(特征)系数被惩罚为 0，从而具备特征选择功能。表 3 给出了 Lasso 在高维小样本数据应用中的优势和不足。

表 3　Lasso 在高维小样本中的优势与不足

Tab. 3　Advantages and disadvantages of Lasso

优点	缺点
具有 Oracle Property 特性，求解特征选择问题时保证了模型的相合性和参数估计渐进正态性	1) 有偏估计，选择出的变量与实际上对类标变量有影响的变量并非完全一致(原因在于不仅对噪声、冗余等回归系数进行压缩，对全部变量对应的回归系数都进行压缩)； 2) 随着数据维度剧增，Lasso 函数值或梯度值计算复杂度和时间复杂度难以满足需求； 3) 由于 Lasso 凸优化，样本与模型的拟合程度较高易出现过拟合问题

为了解决 Lasso 的有偏估计问题，自适应 Lasso、松弛 Lasso、SCAD(Smoothly Clipped Absolute Deviation)模型、MCP(Minimax Concave Penalty)等模型[42-43]相继被提出。

由于缺少高效的求解算法，Lasso 在高维小样本特征选择研究中没有广泛流行，最小角回归(Least Angle Regression，LAR)算法[44]的出现有力促进了 Lasso 在高维小样本数据中的应用，后续研究人员从降低计算复杂度及避免过拟合等角度对 Lasso 进行了改进，如采用迭代思想的 GSIL(Gene Selection based on Iterative Lasso)，采用序贯思想的 SLasso(Sequential Lasso)以及采用有监督组 Lasso(Supervised Group Lasso，SGLasso)等。表 4 给出了部分针对 Lasso 缺点的改进方法及其基本思想。

1.4　Ensemble(集成法)

集成学习是使用一系列特征选择方法进行学习，并使用某种规则把各个学习结果进行整合从而获得比单个特征选择方法更好学习效果的高维小样本特征选择方法。在某些高维小样本特征选择问题研究中，学者多采用这类方法提升特征选择算法的稳定性，如 Li 等[53]提出一种新的特征选择方法，采用重抽样技

术把数据扰动，生成几个不同的训练集和测试集，反复调用递归决策树，并把分类错误率作为评价指标来选择特征；Dutkowski等[53]将不同特征选择算法用于基因选择，并通过优化策略来整合各个算法得到的结果，形成最终的特征子集；Saeys等[54]、Abeel等[55]将Bagging思想用于集成特征选择，Saeys采用融合多个特征选择算法结果的方式来完成特征子集的集成，而Abeel则通过样本重取样技术生成多个特征子集，并在若干高维小样本数据集上与SVM-RFE方法进行对比，取得较好效果。

表4 针对Lasso的改进算法

Tab. 4 Improved algorithms for Lasso

解决问题	算法描述	应用
降低计算复杂度同时避免过拟合	将训练集按维数随机(或均匀)分为K个，分别进行计算，时间复杂度降为K的二次幂或更低，维数与样本数相对平衡避免过拟合；合并所有子集重新进行特征选择	K-split Lasso[45] 均分式Lasso[46]
	采取序列前向搜索，将数据集按特征分为K个，结合迭代优化K次，把每次的最优解与下一个子集合并，得到最优解	GSIL[27] 迭代式Lasso[47]
	与二范式约束条件相结合	弹性网Lasso[48]
其他	将序贯算法思想与Lasso结合，更加稳定且伸缩性更好	SLasso[49-50]
	采用有监督组Lasso，K-means对特征聚类，在每个簇上进行用Lasso和组Lasso分别选择簇内重要的基因和重要的簇	SGLasso[51]

1.5 特征选择方法总结

高维小样本特征选择难点在于在有限的样本空间内尽量剔除冗余特征及噪声数据、尽量选择分辨能力更好的特征、保证算法的稳定性，同时还需要避免过拟合及综合考虑时空性能。表5给出了四类特征选择方法在高维小样本应用中的优缺点对比。

表5 四类特征选择方法优缺点对比

Tab. 5 Advantages and disadvantages of four feature selection categories

算法名称	优点	缺点
Filter	时间性能较好 伸缩性较好 可解释性好	分类精度一般 不能完全去冗余
Wrapper	特征分辨力好 分类正确率较高 能度量特征间相互关系	时空性能差 易过拟合 可解释性差 计算复杂度高
Embedded	不易出现过拟合 去除冗余特征 特征分辨力好	稳定性一般 计算复杂度较高
Ensemble	稳定性好 不易出现过拟合	时间性能一般

在实际应用中，SVM和以Lasso为代表的正则化方法是比较常见的方法，特别是SVM方法由于其产生的分类器结构简单，用到的样本信息很少，受到了更多的关注，在一些文献中也有将两者结合共同进行特征选择的研究[25]。在近期的文献中，已较少见到Filter和Wrapper单独作为高维小样本的特征选择模型，常见的多将其应用于两阶段的特征选择，如使用计算复杂度较低的Filter方法去除冗余与噪声特征，选择与类标签相关度较好的特征构成新的特征集合，再用Wrapper方法(如SVM、GA、随机森林)在新的特征集合上去选择分辨力非常强的特征子集，可以实现很好的降维效果[56-58]，而集成法则更多用于需要算法具备较好稳定性的应用中，对原本就稳定的特征选择方法，集成后效果并不明显。

2 挑战与研究展望

2.1 面临的挑战

1) 数据不一致，各算法很难直接比较。

高维小样本特征选择多源自数据驱动，以基因微阵列数据为例，各研究人员使用的数据来源不尽相同，甚至有同名数据集但内容不相同的情况[59]，这使得后续研究人员很难在统一的数据标准下重现目标算法来进行结果的比较。

2) 分类正确率是否仍然是唯一重要指标。

传统特征选择方法有很多评价指标，在高维小样本环境中，有相当一部分文献在作算法对比分析时，仅仅进行了分类精度的度量，较少涉及算法稳定性、结果可解释性等度量。

除此之外，面对不平衡高维小样本数据，特别是在一些代价敏感学习问题中，降低决策风险、减小平均误分类代价和提高分类可靠性显得尤为重要，分类正确率是否依然作为特征选择唯一的重要的评判标准还应具体问题具体对待。

3) 复杂且超高维数据带来巨大压力。

目前主流的特征选择方法SVM及Lasso等在处理连续型数据时具有较好的优势，随着大数据技术的不断发展，数十万维且混合了离散和连续型数值的数据集将有可能成为高维小样本领域最常见的对象，个别数据甚至可能超过千万维，复杂且超高维数据的到来给未来特征选择方法不但在时间和空间复杂度方面提出了巨大挑战，更在算法本身的设计方面提出了新的要求。

4) 缺少国内可信数据源。

目前学界使用的大多为国外机构提供的分析数据，缺少国内公开的权威数据源，特别是基因微阵列数据及人像识别等领域，据此分析的结果得出的结论(如哪些基因对疾病诊断有帮助)很难在实践中得到验证，各算法得出的最终结果的可解释性受到一定影响。

5) 缺少对特征选择结果的解释分析。

大多数文献采取与其他文献中算法进行量化的对比分析来说明自己算法的优势，很少有文献针对特征选择的最终结果，即最终选择的特征集合的物理属性进行深入分析，仅通过文献的阅读很难了解这些被选择出的属性是否可以用于指导实践，特征选择结果的可解释性无从考证。

2.2 未来研究展望

随着大数据时代的不断演变，高维数据的价值越来越凸显，高维小样本特征选择主要面临计算复杂度、时间复杂度和复杂数据类型等问题，结合文献调研及趋势分析，本文认为面向高维小样本的特征选择方法还可以在以下几个方面取得新

的进展:

1) 增量式学习算法(Incremental Learning) 。

在未来超高维小样本数据面前,常规的集中式特征选择方法可能很难满足时空性能的需求,甚至无法得到令人满意的特征选择结果,增量式学习为超高维小样本数据的特征选择问题提供了一种思路,增量式不仅指数量的逐渐递增,更多的是指特征数量的逐渐递增。目前,已有工作开展了相关研究,形成了被称为"Online Feature Selection"的研究热点。

2) 非连续型数据的处理。

当前高维小样本特征选择所面对的数据对象很大一部分是连续型数据,特别是基因微阵列数据都是细胞内 mRNA 的相对或绝对数量来表示的连续型数据,相应的,大多数文献都选择了 SVM 或 Lasso 等可以直接处理连续型数据的特征选择方法,就本文所调研范围,少有文献去深入分析高维小样本连续型数据的离散化问题。因此,对非连续型数据的特征选择及与连续型数据特征选择方法进行深度对比分析具有一定研究价值。

3) 算法的稳定性(Stability) 与可伸缩性(Scalability) 。

Lasso 等正则化的稀疏模型存在特征选择的不一致性,即模型稀疏化后的不稳定性,同样的问题也可能出现在其他几类主流的特征选择方法中,由于高维小样本数据的特殊性,今后的研究可能不单纯只进行分类正确率及时空性能的比较,还需要考虑算法本身的稳定性,这样选择的结果才更加容易被接受; 同时,随着数据维数的急剧增加,算法的可伸缩性也是一个重要的指标,随着数据量和维数的增加,算法的性能不可出现显著下降。

4) 传统统计学算法焕发新生。

正则化方法产生于 1955 年,Lasso 也是 20 年前的方法,这些传统的方法在高维小样本等新的应用环境里重新获得了发展,这也提示传统的统计学中的变量选择方法是否值得去重新梳理,特别是一些线性计算,空间复杂度低的方法,挖掘和改进使之与当前高维小样本应用能够结合,从而丰富高维小样本的特征选择方法。

5) 多阶段的混合式特征选择。

在高维小样本的应用中,通常伴有高维不平衡、高维稀疏等现象,没有一种通用的方法可以应对所有的高维小样本分类应用问题,采用混合的多阶段的特征选择方法,可以有效去除不相关及冗余特征,如使用重抽样技术扩大样本规模,使用 Filter 去除不相关属性,再使用 Embedded 等方法去除冗余属性等。

6) 可信数据源的构建。

随着政府开放数据带来的利好,国内会有更多专业机构参与国内数据的采集整理工作,如果能根据权威公开数据构建高维小样本数据开放共享平台,让更多研究人员参与其中,可以更好地发挥数据本身的价值。

参考文献 (References)

[1] ESPEZUA S, VILLANUEVA E, MACIEL C D, et al. A projection pursuit framework for supervised dimension reduction of high dimensional small sample datasets [J]. Neurocomputing, 2015, 149(PB): 767 – 776.

[2] LAZAR C, TAMINAU J, MEGANCK S, et al. A survey on filter techniques for feature selection in gene expression microarray analysis [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(4): 1106 – 1119.

[3] TAO H, BAUSCH C, RICHMOND C, et al. Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media [J]. Journal of Bacteriology, 1999, 181(20): 6425 – 6440.

[4] KERR M K, MARTIN M, CHURCHILL G A. Analysis of variance for gene expression microarray data [J]. Journal of Computational Biology, 2000, 7(6): 819 – 837.

[5] THOMAS J G, OLSON J M, TAPSCOTT S J, et al. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles [J]. Genome Research, 2001, 11(7): 1227 – 1236.

[6] EFRON B, TIBSHIRANI R, STOREY J D, et al. Empirical Bayes analysis of a microarray experiment [J]. Journal of the American Statistical Association, 2001, 96(456): 1151 – 1160.

[7] LONG A D, MANGALAM H J, CHAN B Y, et al. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework [J]. Journal of Biological Chemistry, 2001, 276(23): 19937 – 19944.

[8] BALDI P, LONG A D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes [J]. Bioinformatics, 2001, 17(6): 509 – 519.

[9] PARZEN E. On estimation of a probability density function and mode [J]. The Annals of Mathematical Statistics, 1962, 33(3): 1065 – 1076.

[10] WILINSKI A, OSOWSKI S, SIWEK K. Gene selection for cancer classification through ensemble of methods [C] // Proceedings of the 9th International Conference on Adaptive and Natural Computing Algorithms. Berlin: Springer, 2009: 507 – 516.

[11] STEUER R, KURTHS J, DAUB C O, et al. The mutual information: detecting and evaluating dependencies between variables [J]. Bioinformatics, 2002, 18(Suppl. 2): S231 – S240.

[12] LIU X, KRISHNAN A, MONDRY A. An entropy-based gene selection method for cancer classification using microarray data [J]. BMC Bioinformatics, 2005, 6(1): 1 – 14.

[13] CHUANG L Y, KE C H, CHANG H W, et al. A two-stage feature selection method for gene expression data [J]. Omics: a Journal of Integrative Biology, 2009, 13(2): 127 – 137.

[14] GOLUB T R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring [J]. Brain Research, 1999, 501(2): 205 – 214.

[15] 李颖新,李建更,阮晓钢. 肿瘤基因表达谱分类特征基因选取问题及分析方法研究 [J]. 计算机学报,2006,29(2): 324 – 330. (LI Y X, LI J G, RUAN X G. Study of informative gene selection for tissue classification based on tumor gene expression profiles [J]. Chinese Journal of Computers, 2006, 29(2): 324 – 330.)

[16] 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.

[17] PARK P J, PAGANO M, BONETTI M. A nonparametric scoring algorithm for identifying informative genes from microarray data [EB/OL]. [2016-12-17]. http://xueshu.baidu.com/s?wd=paperuri%3A%286c6a741e996db71f799147979ac19d70%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fdx.doi.org%2F10.1142%2F9789814447362_0006&ie=utf-8&sc_us=5571940567161427371.

[18] CHENG Q, ZHOU H, CHENG J. The Fisher-Markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data [J]. IEEE Transactions on Pattern Analysis and Machine intelligence,

2011,33(6):1217-1233.
[19] WANG Y, TETKO I V, HALL M A, et al. Gene selection from microarray data for cancer classification—a machine learning approach [J]. Computational Biology & Chemistry, 2005, 29(1): 37-46.
[20] DING C, PENG H. Minimum redundancy feature selection from microarray gene expression data [J]. Journal of Bioinformatics and Computational Biology, 2005, 3(2): 185-205.
[21] XING E P, JORDAN M I, KARP R M. Feature selection for high-dimensional genomic microarray data [C]// Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001: 601-608.
[22] HIRA Z M, AGILLIES D F. A review of feature selection and feature extraction methods applied on microarray data [J]. Advances in Bioinformatics, 2015, 2015: Article ID 198363.
[23] LI L, WEINBERG C R, DARDEN T A, et al. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method [J]. Bioinformatics, 2001, 17(12): 1131-1142.
[24] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods [J]. Computers & Electrical Engineering, 2014, 40(1): 16-28.
[25] XIA X L, XING H, LIU X. Analyzing kernel matrices for the identification of differentially expressed genes [J]. PLOS ONE, 2013, 8(12): e81683.
[26] OSAREH A, SHADGAR B. Machine learning techniques to diagnose breast cancer [C]// Proceedings of the 2010 5th International Symposium on Health Informatics and Bioinformatics. Piscataway, NJ: IEEE, 2010: 114-120.
[27] 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: 15, 35-52. (ZHANG J. Classification and feature selection on high-dimensional and small-sampling data [D]. Hefei: Hefei University of Technology, 2014: 15, 35-52.)
[28] SUN Y, BABBS C F, DELP E J. A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm [C]// Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society. Piscataway, NJ: IEEE, 2006: 6532-6535.
[29] NAKARIYAKUL S, CASASENT D P. An improvement on floating search algorithms for feature subset selection [J]. Pattern Recognition, 2009, 42(9): 1932-1940.
[30] CHUANG L Y, YANG C H, LI J C, et al. A hybrid BPSO-CGA approach for gene selection and classification of microarray data [J]. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 2012, 19(1): 68-82.
[31] CORDÓN O, DAMAS S, SANTAMARÍA J. Feature-based image registration by means of the CHC evolutionary algorithm [J]. Image & Vision Computing, 2006, 24(5): 525-533.
[32] KAMYAB S, EFTEKHARI M. Feature selection using multimodal optimization techniques [J]. Neurocomputing, 2016, 171(C): 586-597.
[33] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for caner classification using support vector machines [J]. Machine Learning, 2002, 46(1): 389-422.
[34] DING Y, WILKINS D. Improving the performance of SVM-RFE to select genes in microarray data [J]. BMC Bioinformatics, 2006, 7(Suppl 2): S12.
[35] MAO Y, PI D, LIU Y, et al. Accelerated recursive feature elimination based on support vector machine for key variable identification [J]. Chinese Journal of Chemical Engineering, 2006, 14(1): 65-72.
[36] 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 [J]. 计算机学报, 2014, 37(8): 1704-1718. (XIE J Y, XIE W X. Several feature selection algorithms based on the discernibility of a feature subset and support vector machines [J]. Chinese Journal of Computers, 2014, 37(8): 1704-1718.)
[37] 游伟,李树涛,谭明奎. 基于SVM-RFE-SFS的基因选择方法 [J]. 中国生物医学工程学报, 2010, 29(1): 93-99. (YOU W, LI S T, TAN M K. Gene selection method based on SVM-RFE-SFS [J]. Chinese Journal of Biomedical Engineering, 2010, 29(1): 93-99.)
[38] TANG Y, ZHANG Y Q, HUANG Z. FCM-SVM-RFE gene feature selection algorithm for leukemia classification from microarray gene expression data [C]// Proceedings of the 14th IEEE International Conference on Fuzzy Systems. Piscataway, NJ: IEEE, 2005: 97-101.
[39] 吴红霞,吴悦,刘宗田,等. 基于Relief和SVM-RFE的组合式SNP特征选择 [J]. 计算机应用研究, 2012, 29(6): 2074-2077. (WU H X, WU Y, LIU Z T, et al. Combined SNP feature selection based on Relief and SVM-RFE [J]. Application Research of Computers, 2012, 29(6): 2074-2077.)
[40] 林俊,许露,刘龙. 基于SVM-RFE-BPSO算法的特征选择方法 [J]. 小型微型计算机系统, 2015, 36(8): 1865-1868. (LIN J, XU L, LIU L. Feature selection method based on SVM-RFE and particle swarm optimization [J]. Journal of Chinese Computer Systems, 2015, 36(8): 1865-1868.)
[41] TIBSHIRANI R. Regression shrinkage and selection via the Lasso [J]. Journal of the Royal Statistical Society, 1996, 58(1): 267-288.
[42] 刘建伟,崔立鹏,刘泽宇,等. 正则化稀疏模型 [J]. 计算机学报, 2015, 38(7): 1307-1325. (LIU J W, CUI L P, LIU Z Y, et al. Survey on the regularized sparse models [J]. Chinese Journal of Computers. 2015, 38(7): 1307-1325.)
[43] 刘建伟,崔立鹏,罗雄麟. 结构稀疏模型及其算法研究进展 [J]. 计算机科学, 2016, 43(S1): 1-16. (LIU J W, CUI L P, LUO X L. Research and development on structured sparse models and algorithms [J]. Computer Science, 2016, 43(S1): 1-16.)
[44] EFRON B, HASTIE T, JOHNSTONE I, et al. Least angle regression [J]. Annals of Statistics, 2004, 32(2): 407-451.
[45] 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. (ZHANG J, HU X G, ZHANG Y H, et al. K-split Lasso: an effective feature selection method for tumor gene expression data [J]. Journal of Frontiers of Computer Science and Technology, 2012, 6(12): 1136-1143.)
[46] 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. (SHI W F, HU X G, YU K. K-part Lasso based on feature selection algorithm for high-dimensional data [J]. Computer Engineering and Applications, 2012, 48(1): 157-161.)
[47] 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. (SHI W F, HU X G, YU K. Iterative Lasso based on feature selection for high dimensional data [J]. Application Research of Computers, 2011, 28(12): 4463-4466.)
[48] ZOU H, HASTIE T. Regularization and variable selection via the elastic net [J]. Journal of the Royal Statistical Society, 2005, 67(2): 301-320.

(下转第2448页)

进行透明度的调节已经无法满足要求。因此下阶段工作考虑如何使用边捆绑技术对具有相似走向的路径进行捆绑，使界面更整洁清晰。

参考文献（References）

[1] LEI H, XIA J, GUO F, et al. Visual exploration of latent ranking evolutions in time series [J]. Journal of Visualization, 2016, 19(4): 783 - 795.

[2] 姜婷婷，肖卫东，张翀，等. 基于桑基图的时间序列文本可视化方法 [J]. 计算机应用研究，2016，33(9)：2683 - 2687. (JIANG T T, XIAO W D, ZHANG C, et al. Text visualization method for time series based on Sankey diagram [J]. Application Research of Computers, 2016, 33(9): 2683 - 2687.)

[3] BOUALI F, DEVAUX S, VENTURINI G. Visual mining of time series using a tubular visualization [J]. The Visual Computer, 2016, 32(1): 15 - 30.

[4] WEBER M, ALEXA M, MÜLLER W. Visualizing time-series on spirals [C] // Proceedings of the 2001 IEEE Symposium on Information Visualization. Washington, DC: IEEE Computer Society, 2001: 7 - 14.

[5] CARLIS J V, KONSTAN J A. Interactive visualization of serial periodic data [C] // Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology. New York: ACM, 1998: 29 - 38.

[6] CHENG S, JIANG Z, QI Q, et al. The polar parallel coordinates method for time-series data visualization [C] // Proceedings of 2012 4th International Conference on Computational and Information Sciences. Washington, DC: IEEE Computer Society, 2012: 179 - 182.

[7] HEWAGAMAGE K P, HIRAKAWA M, ICHIKAWA T. Interactive visualization of spatiotemporal patterns using spirals on a geographical map [C] // Proceedings of the 1999 IEEE Symposium on Visual languages. Washington, DC: IEEE Computer Society, 1999: 296 - 303.

[8] DRAGICEVIC P, HUOT S. SpiraClock: a continuous and non-intrusive display for upcoming events [C] // Proceedings of the CHI '02 Extended Abstracts on Human Factors in Computing Systems. New York: ACM, 2002: 604 - 605.

[9] TOMINSKI C, SCHUMANN H. Enhanced interactive spiral display [EB/OL]. [2016-11-25]. http://www.informatik.uni-rostock.de/~schumann/papers/2008%2B/tominski_spiral_display.pdf.

[10] BUELENS B. Visual circular analysis of 266 years of sunspot counts [J]. Big Data, 2016, 4(2): 89 - 96.

[11] WALKER J, BORGO R, JONES M W. TimeNotes: a study on effective chart visualization and interaction techniques for time-series data [J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 549 - 558.

This work is supported by National Natural Science Foundation of China (61573292).

YANG Huanhuan, born in 1991, M. S. candidate. Her research interests include data visualization.

LI Tianrui, born in 1969, Ph. D., professor. His research interests include particle calculation and rough set, data mining and knowledge discovery, cloud computing and big data.

CHEN Xindi, born in 1992, M. S. candidate. Her research interests include data visualization.

（上接第2438页）

[49] 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.

[50] CHEN Z H. Sequential Lasso for feature selection with ultra-high dimensional feature space [EB/OL]. [2016-11-25]. http://www.stat.nus.edu.sg/~stachenz/T11-455R1.pdf.

[51] MA S, SONG X, HUANG J. Supervised group Lasso with applications to microarray data analysis [J]. BMC Bioinformatics, 2007, 8(1): 1 - 17.

[52] LI X, RAO S, WANG Y, et al. Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling [J]. Nucleic Acids Research, 2004, 32(9): 2685 - 2694.

[53] DUTKOWSKI J, GAMBIN A. On consensus biomarker selection [J]. BMC Bioinformatics, 2007, 8(Suppl 5): S5.

[54] SAEYS Y, ABEEL T, PEER Y V D. Robust feature selection using ensemble feature selection techniques [C] // Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases, LNCS 5212. Berlin: Springer, 2008: 313 - 325.

[55] ABEEL T, HELLEPUTTE T, VAN DE PEER Y, et al. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods [J]. Bioinformatics, 2010, 26(3): 392 - 398.

[56] WANG Y, MAKEDON F S, FORD J C, et al. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data [J]. Bioinformatics, 2005, 21(8): 1530 - 1537.

[57] AKADI A E, AMINE A, OUARDIGHI A E, et al. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper [J]. Knowledge and Information Systems, 2011, 26(3): 487 - 500.

[58] BERMEJO P, DE LA OSSA L, GÁMEZ J A, et al. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking [J]. Knowledge-Based Systems, 2012, 25(1): 35 - 44.

[59] BOLÓN-CANEDO V, SÁNCHEZ-MAROÑO N, ALONSO-BETANZOS A, et al. A review of microarray datasets and applied feature selection methods [J]. Information Sciences: an International Journal, 2014, 282(5): 111 - 135.

[60] 姚唐龙. 基因表达谱数据挖掘的特征提取方法研究 [D]. 合肥：安徽大学，2015：13 - 19. (YAO T L. Research on feature extraction method of gene expression profiles data mining [D]. Hefei: Anhui University, 2015: 13 - 19.)

This work is partially supported by the National Basic Research Program (973 Program) of China (2016YFC0801406), the National Natural Science Foundation of China (61673152), the Natural Science Foundation of Anhui Province (1408085QF136).

WANG Xiang, born in 1982, Ph. D. candidate. His research interests include data mining, artificial intelligence, intelligence analysis.

HU Xuegang, born in 1962, Ph. D., professor. His research interests include data mining, artificial intelligence, big data analysis.