



基于实例学习和协同子集搜索的特征选择方法

许小媛¹, 黄黎^{1,2}

(1. 江苏开放大学信息与机电工程学院, 江苏 南京 210017;
2. 南京航空航天大学计算机科学与技术学院, 江苏 南京 210016)

摘要: 特征子集搜索是数据挖掘分类任务中一个关键性的难题, 常用的过滤器方法忽略了基因之间的相关性, 此外, 现有的解决方法并不是专门针对处理小样本数据, 因此在特征选择方面表现出了不稳定性。为了解决上述问题, 在实例学习的基础上提出了一种新型的混合封装过滤算法, 并且提出了一种具有封装器评价体系的分类器算法——协同子集搜索 (CSS)。选取几个高维小样本的癌症数据集作为数据来源, 对提出的评价体系进行了实验测试, 结果表明, 该方法在准确性及稳定性方面较其他方法表现更好。

关键词: 特征选择; 混合; 小样本; 分类; 稳定性

中图分类号: TP312

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017122

A feature selection method based on instance learning and cooperative subset search

XU Xiaoyuan¹, HUANG Li^{1,2}

1. College of Information and Mechanical Engineering, Jiangsu Open University, Nanjing 210017, China

2. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract: Feature subset selection is a key problem in such data mining classification tasks. In practice, the filter methods ignore the correlations between genes which are prevalent in gene expression data, additionally, existing methods are not specially conceived to handle the small sample size of the data which is one of the main causes of feature selection instability. In order to deal with these issues, a new hybrid, filter wrapper was proposed, and a cooperative subset search (CSS), was then researched with a classifier algorithm to represent an evaluation system of wrappers. The method was experimentally tested and compared with state-of-the-art algorithms based on several high-dimension allow sample size cancer data sets. Results show that the proposed approach outperforms other methods in terms of accuracy and stability of the selected subset.

Key words: feature selection, hybrid, small sample, classification, stability

收稿日期: 2016-12-28; 修回日期: 2017-03-06

基金项目: 江苏省教育科学“十三五”规划 2016 年度青年专项课题 (No.C-b/2016/03/25); 江苏开放大学“十三五”规划 2016 年度重点研究课题 (No.16SSW-Z-003)

Foundation Items: Youth Project of “13th Five-Year” Plan of Jiangsu Education Science (No.C-b/2016/03/25), Key Project of “13th Five-Year” Plan of Jiangsu Open University (No.16SSW-Z-003)



1 引言

现今, 高产能的生物技术(例如基因芯片和测序方法)可以很容易地测量成千上万个基因。在癌症数据的分类中, 预测模型通常是建立在健康人群和癌症患者的训练数据集上。分类算法确定了基因表达特征和患者标签之间的关系^[1]。在基因表达数据中, 存在成千上万的基因表达, 但典型样本数据不足, 因而多种传统统计方法在应用中出现了严重的问题。其中之一就是分类器的过拟合问题。在训练数据的分类过程中它常常能发挥出非常好乃至完美的性能。但这种完美的表现在新的未标记的数据集中无法体现, 从而导致了应用的有限性。Kohane 等人^[2]认为基因组应用中的小样本量可能与微阵列测定的高成本有关。每个样品涉及成千上万个变量的测量, 所需测量的变量与应用微阵列技术测量的成千上万个基因表达有关。这导致与样本量相比还有大量的需测量的变量数据。对于这样的高维数据, 使用过滤器的方法来测量每个基因和类别标签之间的关系在实践中是非常常见的。然而 Tolosi 和 Lengauer^[3]的实验表明, 过滤器方法忽略了基因之间的相关性, 因为基因的共调节现象在基因数据中是普遍存在的, 结果导致许多冗余的分化基因被纳入, 与此同时有用但弱分化的基因则被忽视。

为了解决上述问题, 本文提出了一种混合特征选择方法, 该方法更有利于从小样本量的数据中得到有用的结果。本文方法在过滤步骤中使用了基于实例的候选特征子集选择, 其中的关键思想是将任意复杂的问题分解为一个集合, 然后通过本地学习的特征相关性在全局找到相关特征。主要目标是对数据集降维, 同时在准确性、特征选择的稳定性和获得子集的大小方面维持良好的性能。

2 特征选择的基本概念

在大多数特征选择技术中, 通常用评估函数对特征子集的分配进行评分, 同时用搜索算法搜索具有高评分的子集。评估函数既可以基于特征与预测的一般相关性度量(过滤器模型), 也可基于特定的预测器(封装器模型)。结合过滤器和封装器的第3类算法被称为混合方法。

2.1 过滤器

过滤器^[4]不依赖于特定类型的预测模型, 仅考虑数据的特征以选择最佳特征子集或通过向每个特征的分配评分来获得特征的排序。该过程在学习过程开始之前就已完成。过滤方法的处理过程很快, 因此在选择高维数据集的特征时非常有用。在已有参考文献中提及了几种过滤器方法, 同时显示了它们在选择最相关特征及提高预测性能方面的有效性。下面描述一些最常见的过滤器方法。

(1) t 检验过滤器

t 检验通常用于特征选择, 将样本特征的得分定义为两类样本平均值与方差的差异率。因此, 每个特征的权重由其计算出的绝对得分给出。

(2) 最小冗余最大相关原则

Peng 等人^[5]提出了最小冗余最大相关(minimal redundancy maximal relation, mRMR)原则。该原则是在样本维数为 m 的 N 个样本和 c 类故障的特征样本空间中选取一个样本维数为 p ($p \leq m$) 的子空间 S 构成最优故障特征空间。这种方法因所选择的特征彼此尽可能不同而在应用上存在限制。

(3) Relief

这种方法是由 Kira 和 Rendell 提出的^[6]。Relief 基于实例学习, 它通过从不同的类中选择特征以区分实例。参考文献[7]将 Relief 的相关性评估标准与边际最大化假设相关联, 这也是该算法在许多应用中具有优越性的原因之一。

过滤器可用作预处理,以减少空间维度并克服过度拟合^[4,5]。当特征的数目变得非常大时,通常选择过滤器模型,因为它在计算上高效快速并且独立于分类算法。然而,考虑到学习算法的预测性能,同时进行特征的选择可能会有很大的益处,增强性能是特征选择的主要目标之一,而过滤器忽略了这些方面,这是其主要缺点。

2.2 封装

结合预测算法的特定偏差和性能来进行最佳特征子集的搜索具有很大的效用。基于此,封装模型使用特定分类器来评估所选特征的质量^[4]。学习算法的性能测量以及诸如交叉验证的重采样算法^[6]用于选择最佳特征子集。给定一个预定义的分类器,典型的封装器模型能基于搜索程序迭代地产生一组特征,然后使用分类器的性能来评估特征,直到达到具有期望质量的特征集。此外,还可以使用广泛的搜索策略,如下所示。

(1) 顺序特征选择

它是使用最广泛的封装技术之一^[4,8]。它通过向前或向后搜索来选择特征的子集,其包括分别添加或去除特征,直到满足给定的终止条件。

(2) 随机特征选择

利用随机封装算法随机搜索下一个特征子集^[9]。可以根据对预测性能的影响,从先前的特征集中一次性添加、移除或替换单个特征或几个特征。

(3) 支持向量机和递归特征消除

尽管在前面的封装器方法中搜索和评估过程是分开的,但是在实际中存在嵌入模型的方法,使用其内部参数将对特征最佳子集的搜索过程构建到分类器中。

封装器通常能为特定类型的模型提供性能最好的特征集,并且能够在考虑分类特征的同时考虑其依赖性。然而,封装器缺乏通用性是其缺点。

不同的学习算法可能导致不同的特征选择结果。此外,封装器在每个候选子集上重复构建学习模型,其运算过程极为耗时,这是该类方法的主要问题。

2.3 混合方法

与封装器模型相关联的计算成本使得当特征的数量过高时封装器不具有可行性,而过滤器的性能不太稳定。混合模型很好地结合了两种方法并克服上述缺陷。混合方法使用过滤器来生成特征的排序列表,基于此排序,学习机使用封装器算法生成和计算具有良好定义的嵌套子集。

在高维和小样本癌症数据诊断的背景下,本文提出使用混合方法,其将微阵列数据集的小样本转换为用于指导特征子集搜索的工具。

3 基于特征选择算法的混合实例

本文混合方法使用的过滤技术来源于以Relief为基础的方法^[6]。因为特征选择的实例为随机选择,因此该方法具有不稳定性。特征的权重可能随实例^[8]而波动,使得特征对数据采样敏感,特别是在存在数据噪声和高维离群值的情况下。然而,本文方法使用所有实例来对特征进行加权,结合特征相关性的本地学习,第1步通过基于实例的学习生成候选特征子集(candidate feature subset, CFS),然后使用组合技术在全局找到相关的特征。这个特征选择的目的不是为某个实例找到好的特征子集,而是根据多重候选子集找到最多的实例。此外,在评估CFS并指导组合过程的第2步中引入包装器方法,旨在利用特定的学习算法来搜索最佳特征子集并实现最佳性能。对于这一步,本文提出一个CSS算法,它是通过了10层交叉验证的最好的特征子集搜索技术。图1中为基于混合实例的协作子集搜索(HIB-CSS)方法。其中的两个步骤将在第3.3、3.4节中详细说明。

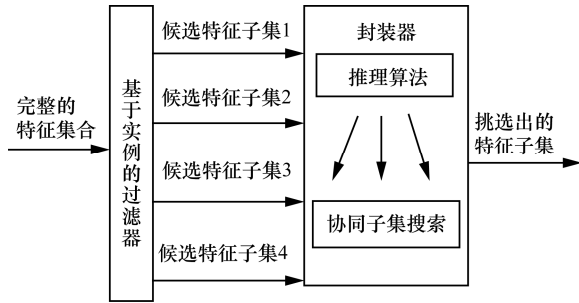


图1 基于实例的协作子集搜索

3.1 筛选步骤：基于实例的候选特征子集选择

令 X 是包含 m 个训练实例的矩阵 $x_i=(x_{i1}, x_{i2}, \dots, x_{id}) \in R^d$, 其中 d 是特征的数量, $y_i=(y_1, \dots, y_m)$ 是 m 个实例的类标签的向量。令 A 为如下的一系列特征 $a_j=(a_{1j}, \dots, a_{mj})$ 的集合, 其中 $d \gg m$ 。

在最优特征子集选择的预处理步骤中, 将特征空间减少到 m 个候选子集。训练数据的每个实例都是一个专家, 它在实例特征加权技术的基础上提出了一个候选特征子集。给定一个距离函数, 为每个特征 a_j 找到每个样本 x_i 的两个近邻, 一个来自同一类 (称为最近命中或 NH), 另一个来自不同的类 (称为最近缺失或 NM)。然后使用距离函数计算 x_{ij} 的边界值:

$$M(x_{ij})=d(x_{ij}, NM(x_{ij}))-d(x_{ij}, NH(x_{ij})) \quad (1)$$

本文使用曼哈顿距离来定义样本的边缘和最近邻, 也可使用其他标准函数。该权重定义曾在众所周知的 Relief 算法中得到使用 (欧几里得距离), 并用于特征选择^[7]。有人认为, 在 Relief 算法的评估中使用两个不同的指标——欧氏或曼哈顿^[8], 在结果上未发现显著性差异。通过参考文献发现本文算法中的距离函数与 Relief 算法几乎一样, 因此有理由相信本文算法的性能不会受距离函数选择的影响。然后将这些得分标准化为单位长度, 使得分的完整向量具有一个长度。通过将每个分数除以分数向量的欧几里德长度来完成上述步骤。因此, 获得 x_i 的加权特征空间。投影在每个特征 a_j 上的权重是 $w_{aj}=w_{1j}, \dots, w_{mj}$ 。标准化特征权重矩阵 W , 见表 1。

表1 特征权重矩阵

样本	特征值 1	特征值 2	...	特征值 d
样本 1	$w_{1,1}$	$w_{1,2}$	...	$w_{1,d}$
样本 2	$w_{2,1}$	$w_{2,2}$	...	$w_{2,d}$
样本 3	$w_{3,1}$	$w_{3,2}$	...	$w_{3,d}$
...	...	...	...	...
样本 m	$w_{m,1}$	$w_{m,2}$	...	$w_{m,d}$

然后, 在每个实例空间中的特征以实例的权重为基础进行排名。注意, 实例可以提供与另一实例不同的特征排序。因此, 根据所考虑的实例, 特征 a_j 可以具有不同的排名。在基于实例的特征排序步骤之后, 从每个实例排名靠前的特征中选择基数为 n 的候选子集。这个预处理步骤会形成基数为 n 的 m 个候选特征子集 $CFS_{all}=\{CFS_1, CFS_2, \dots, CFS_m\}$, 本实验对基数范围在 1~20 的 10 个候选特征子集 n 进行测定, 以探寻算法的最佳性能。

要注意的是, HIB-CSS 方法假定如果一个特征是好的, 即使对于某些实例它没有被选择, 那么它会因其他实例的选择而出现在排名靠前的特征中。此外, 如果这个特征足够好, 它将影响 CFS 的分类结果, 并增加其出现在最终子集中的概率。

3.2 封装步骤：协同子集搜索

在第 2 个步骤中, CFS 被集成在最佳特征子集的搜索过程中, 其中子集搜索和分类器算法组成了封装器的评估系统。该封装器方法基于 CSS, 在并行而非迭代方式下以训练实例的特征选择对分类性能的影响来对特征进行组合。在过滤器步骤中获得的特征权重作为封装器的输入。 w_{ij} 的值取决于特征 a_j 是否出现在候选子集 CFS_i 中。因此, 特征权重将采用:

$$w_{k,j} = \begin{cases} w_{i,j}, & w_{i,j} \in CFS_i \\ 0, & \text{其他} \end{cases} \quad (2)$$

由于 CSS 过程包括评估 CFS_s 的封装步骤, 需要对其进行数据测试。因此, 初始交叉验证的每个迭代 k 被用于评估整体特征选择方法的性能, 另一交叉验证仅应用于该迭代的训练数据, 并仅在封装器步骤进

行验证。这意味着从初始交叉验证的迭代 k 的训练数据中获得其子样本——封装测试数据。相应的初始测试数据仅用于整个特征选择模型评估,一旦被用于构建迭代 k , 则不进行偏置性能分析。

分类器的选择影响到了封装算法的效率,对两种协同子集搜索中的分类算法 SVM 和 KNN 进行了比较,见表 2。在 CSS 的过程中,使用 10 层交叉验证来对每个 CFS_s 投射的分类器算法进行训练,并且分类误差 β_i 及相关训练数据由该交叉验证进行计算。在这种分类之后,获得 FS_{Good} , 其包含较优 CFS_s 中的某些特征,而 FS_{Bad} 包含较差 CFS_s 中的某些特征。它们的组成特征被收集在名为 FS_{Union} 的单特征子集中,它们的特征权重基于对应的错误率并同步更新。基于所计算的权重,对 FS_{Union} 中的特征进行排序,并且从中选择 S 个最佳特征。此步骤选定的子集将在下列两个步骤中更新:从 FS_{Union} 中提取 FS_{Bad} 的特征;添加 FS_{Good} 的特征到 FS_{Union} 中,返回特征子集 FS_{Union} 作为最佳特征子集。注意,一旦获得了最终特征子集,就可以删去预选步骤的特征子集 S , 并且用几个候选特征子集进行代替。

4 实验研究

本节阐述了提出的混合特征选择方法的实验设置和结果以及该方法与 7 种现有方法的比较结果。同时将混合算法与 7 种现有方法应用于几个高维数据集。

4.1 数据集

实验在 7 个高维小尺寸微阵列数据集和大规

模数据集上进行。7 个高维数据的资料为二分类资料,要完成的任务是肿瘤的诊断。样本为 Shipp 等人^[10]提出的弥漫性大 B 细胞 (diffuse large B-cell, DLBCL), 其分类任务是组织类型的预测,其中基因用于区分 DLBCL 组织和滤泡性淋巴瘤。Dyrskjot 等人^[11]描述的膀胱癌数据处理的任务是使用微阵列对膀胱 (bladder) 肿瘤进行临床分类。本文还考虑另一个淋巴瘤 (lymphoma) 数据集,其任务是通过微阵列技术测量的基因表达来区分两种类型的淋巴瘤。此数据集包含数值属性的缺失,本文使用 Troyanskaya 等人^[12]提出的 KNN 估算方法来替换它们。由 Singh 等人^[13]描述的前列腺 (prostate) 数据集包含 102 个样品 12 600 个基因的表达,包括前列腺肿瘤样本和正常样本。另一个数据集是 Van't Veer 等人^[14]使用的乳腺癌数据集。同时也考虑了 Pomeroy 等人^[15]提出的中枢神经系统 (CNS) 数据集。它是一个以基因表达为基础与中枢神经系统胚胎肿瘤预测有关的大数据集。还分析了 Gordon 等人^[16]使用的恶性胸膜间 (breast) 皮瘤和肺腺癌基因表达数据库 (lung), 其目的是区分恶性胸膜间皮瘤和肺癌。还对来自 UCI 存储库的大规模数据库——Gisette^[17]进行分析,其样本来源于手写的图片,以区分两个易混淆的手写数字——4 和 9。由于实例的数量高于特征集大小,该数据集与既往实验采用的数据集相比有很大的不同,该数据集允许在大规模数据集上使用 HIB-CSS 方法,即使该方法最初是专为小样本数据集而设计的。对 Gisette 数据库的大量实例进行分析需要太长的执

表 2 数据集的协同子集搜索结果

数据集	协同子集搜索 (SVM 算法)			协同子集搜索 (KNN 算法)		
	候选特征子集	终选特征子集	最小分类误差	候选特征子集	终选特征子集	最小分类误差
DLBCL	10	68	0.012 5	5	28	0.050 0
膀胱癌	9	50	0.033 3	9	17	0.033 3
淋巴瘤	7	28	0.000 0	5	7	0.000 0
前列腺癌	9	69	0.050 0	2	25	0.099 1
胸膜癌	10	92	0.268 3	9	100	0.288 9
CNS	8	36	0.209 5	1	1	0.231 9
肺癌	5	25	0.005 3	5	129	0.000 0
UCI	14	200	0.063 7	5	189	0.090 8



行时间,为解决该问题本文使用了在混合方法过滤步骤中所使用的用于生成 100 个 CFS 实例的随机选择方法。表 3 总结了纳入研究的数据集的特征。

4.2 性能指标

本文使用 $K=10$ 的 K -fold 分层交叉验证来预测 KNN 算法的分类性能,并测试所有特征选择方法在 8 个数据集上的稳定性。整个特征选择过程建立在除了第 K 个之外的所有子集上。模型 K 的性能在第 K 个子集上被评估,该评估不用于对该模型的评估。需要注意的是,如果对所有数据执行特征选择,然后交叉验证,则交叉验证过程的测试数据也用于选择特征,并且这是偏差性能分析。但是,如果采用正确的操作,并在每个折叠中执行特征选择,则在折叠中使用的特征选择中不再存在其他任何有关信息。然后求第 K 个独立测量的性能平均值。还使用此设置来测量在多重算法运行中获得的特征子集的稳定性。Li 等人^[18]认为,如果交叉验证特征集在不同的折叠之间存在差异,它表明特征选择存在不稳定性。还评估了所获得的最终选择特征子集(selected feature subset, SFS)的基数,用于将本文提出的混合方法与现有特征选择方法进行评估比较。

(1) 分类性能

分类器的错误分类误差(misclassification error, MCE)定义为所有分类实例中错误分类实例的比例。此指标很重要,并且一直用于评估分类任务算法的特征选择。

(2) 稳定性

特征选择算法的稳定性是特征偏好的顽健性,它用于区分来源于相同分布的训练集的差异^[19]。测量稳定性需要与特征优选项进行相似性测量,将测量 K 个样本集共有相同特征的程度。

(3) 共同特征

使用交叉验证来产生如前所述的集合。参考文献[20]提出了以下稳定性指标:

$$Stab(S_1, \dots, S_k) = \frac{1}{K(K-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^K (|S_i \cap S_j| - \frac{s^2}{d}) / (S - \frac{s^2}{d}) \quad (2)$$

其中, d 是特征的总数,并且 S_i, S_j 是以训练样本的不同部分所构建的两个特征集。比率 s^2/d 校正了在两个集合中选择的公共特征^[18]的偶然偏差。该指标满足 $-1 < Stab < 1$, 并且其值越大,各组中共同特征的数量也越多。负值意味着公共特征的特征集的产生是偶然导致。

4.3 算法的性能

本文对提出的混合方法进行试验以寻找最佳性能,设定基数范围为 1~15,初始 CFS 为 10,并使用基数特征多达 100 个的 SFS 来测试 HIB-CSS 的表现。对于在封装步骤中使用 KNN 或者 SVM 作为分类器的两种方法,分别记录其各自的 MCE 和对应的 SFS 基数。表 2 显示了本文提出的混合方法在数据集上的结果。获得的最小 MCE (min MCE),并报告了相应的 CFS 及 SFS 基数。

HIB-CSS 在膀胱和淋巴瘤数据集上表现出完全相同的分类性能。对于其他数据集,使用 HIB-CSS (SVM) 获得的预测性能在大多数情况下表现得更好。HIB-CSS (KNN) 在两个数据集(淋巴瘤和肺癌)上以 0% (100%准确度) 获得了最佳 MCE。在 5 个数据集中使用 HIB-CSS (KNN) 算法,所选定的特征中 SFS 基数范围为 1~28。对于其他 3 个数据集,其基数 ≥ 100 。然而,除肺癌数据集外,在封装阶段,使用 SVM 的算法获得最佳性能所需的特征的数量高于使用 KNN。在肺癌数据集除使用 SVM 的算法获得最佳性能所需的特

表 3 数据集的特征

数据集	DLBCL	膀胱癌	淋巴瘤	前列腺癌	胸腺癌	CNS	肺癌	UCI
样本数/个	77	31	45	102	97	60	181	7 000
特征数/个	7 028	3 035	4 025	12 600	24 482	7 129	12 533	5 000

征的数量与使用 KNN 比较, 其数量接近或更小。如果优先选择小样本量的 SFS, 则可以牺牲一些小的分类性能。例如, 使用 HIB-CSS (KNN) 的肺癌数据集的结果显示只选择 11 个特征子集, CFS 基数为 1 时得到的 MCE 为 0.005 3, 该方法与获得了 25 个特征的 HIB-CSS (SVM) 性能相同。

因此, 在这种情况下, 可以根据其性能优先级来选择要使用的 SFS 基数, 即选择最优的 MCE 或最优 SFS 基数。需要注意的是, 对于 CNS 数据库只有通过 HIB-CSS (KNN) 特征选择并以最小 CFS 基数为 1 的初始设置才得到了最佳 MCE。另外, 在 Gisette 的分类中得到了最高的 SFS 基数, 而 Gisette 是大规模数据集。SVM 和 KNN 作为在 HIB-CSS 方法为在封装步骤中使用的两种分类器, 对比研究表明与 KNN 分类器相比, SVM 能产生类似或更好的分类性能, 但该结论需要更多的试验及特征进行验证。

4.4 与其他算法的比较

将本文提出的方法和 7 个特征选择方法进行比较, 其中这 7 个方法有: Relief、t 检验和 mRMR 算法 (属于过滤器)、随机化算法 (属于一种封装器)、SVM-RFE^[12]和特征生成机 (FGM) 算法 (属于嵌入式方法) 以及一种顺序混合算法。FGM 算法^[21]迭代地生成和学习一组带有信息且稀疏的特征子集, 这些子集向 SVM 输入特征, 随后再使用多核学习算法进行整合。

在表 4 中给出了最佳分类结果、相应特征子集的稳定性。为了更加清楚地显示结果, 突出显示了最佳分类表现 (MCE) 以及各数据集的最佳稳定性。具有最佳优化的 MCE 稳定性组合才是一个好的特征选择算法。

结果表明, 不论是使用 SVM 还是 KNN, HIB-CSS 都是优化 MCE 结果最好的算法。在测试的 8 个数据集中, HIB-CSS (SVM) 在其中 6 个数据集的测试结果达到最优。然而, 在其他数据集的情况就不是这样了, 可以从中看出

HIB-CSS (KNN) 对于 CNS 和肺部数据集来说是稳定的, 但 HIB-CSS (SVM) 则不是这样。因此, 如果想同时优化分类和稳定性, HIB-CSS (KNN) 在此处表现得更好。由表 4 数据可以看到, 这一算法在 4 组数据集 (DLBCL、乳腺、CNS 和肺癌) 中给出了最佳的 MCE 稳定性表现。对于 Gisette 大规模数据来说, HIB-CSS 算法能够给出与最佳相比良好的分类表现, 尤其是 HIB-CSS (SVM) 达到了 6.37%。但是算法的稳定性较差, 这可以解释为随机选择了用于生成 CFS 的案例数量较少。为此, 使用更多的案例可以同时增加稳定性和分类结果, 但同时也会消耗很多执行时间。

mRMR 过滤器在淋巴瘤数据集中展示出了完美的分类表现, 其优异性等同于 HIB-CSS 和 Hyb-Seq 算法以及一个对前列腺癌数据集的最佳 MCE 稳定性表现。mRMR 分类和稳定性表现尤其是在大规模数据集中取得良好的表现, 可以从中看出 4.58% 的最佳分类表现和 95% 的稳定性, 这与其他方法拉开了显著的差距。对于其他数据集来说, mRMR 和 SVM-RFE 方法的分类结果十分具有竞争力。当 SVM-RFE 在某些数据集中给出较好的结果时, mRMR 就在其他一些数据集表现良好。

对于膀胱和淋巴瘤数据集, Hyb-Seq 算法表现出了最佳的 MCE-稳定性组合。它给出的前列腺癌数据集的 MCE 最小, 具有 34% 的低稳定性 (34%), 在乳腺癌数据集的表现: 52% 的 MCE 和 100% 的稳定性。因此, 这些案例中权衡的 MCE 稳定性难以使人满意。

通常, FGM 或 t 检验方法可以得出最佳的稳定性。对于膀胱癌、淋巴瘤和肺癌数据集来说, t 检验可以给出良好而均衡的分类—稳定性。然而, t 检验在其他数据集中给出的 MCE 结果最差 (DLBCL、乳腺癌和 CNS 癌数据集分别为 14%、52% 和 36%)。在基于 SVM 算法的 FGM 中可以观察到同样的表现, 可以从中看出获得的稳定性常



表 4 几种算法的最佳分类效果和稳定性比较

数据集	HIB-CSS (SVM)	HIB-CSS (KNN)	Hyb-Seq	随机化 算法	SVM-RFE	FGM	t 检验	Relief	mRMR
DLBCL	0.012 5	0.050 0	0.103 9	0.051 9	0.051 9	0.064 9	0.1429	0.064 9	0.077 9
	0.718 5	0.689 3	0.207 8	0.070 9	0.510 2	0.632 6	0.879 1	0.633 5	0.565 7
膀胱癌	0.033 3	0.033 3	0.322 3	0.064 4	0.032 3	0.096 8	0.032 3	0.064 5	0.064 5
	0.544 5	0.592 5	0.738 9	0.066 7	0.437 2	0.499 2	0.687 7	0.465 0	0.403 9
淋巴瘤	0.000 0	0.000 0	0.000 0	0.066 7	0.022 2	0.155 6	0.022 2	0.022 2	0.000 0
	0.512 5	0.233 5	0.715 8	0.068 7	0.435 0	0.730 4	0.723 4	0.429 1	0.595 1
前列腺癌	0.050 0	0.099 1	0.068 5	0.088 2	0.127 5	0.480 4	0.107 8	0.215 7	0.078 4
	0.627 8	0.655 4	0.342 3	0.021 6	0.586 3	0.911 1	0.794 4	0.627 5	0.709 6
胸膜癌	0.268 2	0.288 9	0.525 9	0.268 0	0.299 0	0.505 2	0.525 8	0.443 3	0.278 4
	0.488 4	0.471 8	1.000 0	0.016 8	0.380 6	1.000 0	1.000 0	0.362 8	0.335 5
CNS	0.209 5	0.231 9	0.316 7	0.333 3	0.300 0	0.500 0	0.366 7	0.366 7	0.300 0
	0.208 6	0.622 2	0.121 7	0.044 7	0.491 1	0.964 4	0.397 0	0.497 4	0.379 2
肺癌	0.005 3	0.000 0	0.027 5	0.055 2	0.005 5	0.071 8	0.016 6	0.011 0	0.005 5
	0.495 8	0.880 4	0.121 7	0.011 2	0.483 1	1.000 0	0.836 8	0.840 9	0.661 0
UCI	0.063 8	0.090 8	0.096 7	0.133 0	0.079 9	0.047 6	0.096 3	0.471 1	0.045 8
	0.233 3	0.101 0	0.156 7	0.020 2	0.403 2	0.506 3	0.767 1	0.589 1	0.951 2

常伴随着很差的分类表现。这影响了 FGM 和 t 检验在分类表现方面的可靠性。值得注意的是, 本文实验也证明了 FGM 在大规模数据集中良好的分类表现, 因为 FGM 在 mRMR 之后也展现出了在 Gisette 数据集具有强有力的表现 (4.76%)。随机化的算法在所有的数据集中展示了非常差的稳定性表现, 它仅在一个数据集 (乳腺癌) 中取得了最佳的 MCE。最后, Relief 过滤器在 DLBCL 和肺癌数据集中表现良好, 但在其他数据集中表现得不是特别好。

总之, HIB-CSS 方法赢得了 MCE 稳定性优化的挑战。除了不稳定的随机化算法之外, 其他所有算法的表现因数据集而异, 它们难以同时表现出良好的分类或良好的稳定性, 这使得它们的可靠性不如 HIB-CSS 算法。

5 结束语

为了避免数据过度拟合, 当特征数目相对样本量过大时, 需要进行特征选择。过滤器特征选择方法因其运行速度快非常适合于这种场景。然而, 它忽略了特征之间的相关性及其与学习算法的相互作用, 因此在应用中可能产生较保守的分类性能。

而封装器以感应算法的偏差来选择特征, 该方法通常效果更好。然而, 在大型数据集上封装器方法因其计算量太大而得不到应用。本文提出一种基于协作子集搜索的新型混合方法——HIB-CSS。该方法在其过滤步骤中使用实例学习, 其主要目标是通过减少封装器评估的数量来加速特征子集的选择过程, 同时在子集的精度、稳定性和大小方面维持良好的性能。该方法的主要挑战点是, 将小样本量的问题转换为一种工具, 此种工具允许仅需选择其中几个变量子集进行分析, 因为 CFS 的数量是实例的数量, 因此, 封装器评估的对象数量显著减少。基于 7 个高维低样本数据集和一个大型数据集, 对提出的方法进行实验测试, 结果表明, 与不同的特征选择方法相比, HIB-CSS 在 MCE 稳定性方面表现最佳。因为过滤器计算成本低是高维数据集首选的要素, 本文期望通过使用小样本数据创建一个高效的过滤器, 以处理特征冗余的问题。

参考文献:

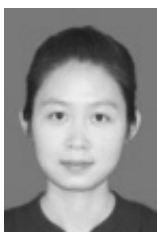
- [1] 刘德山, 范雅惠, 闫德勤, 等. 一种新的去无关基因肿瘤样本分类方法[J]. 辽宁师范大学学报(自然科学版), 2015, 38(1): 41-46.
LIU D S, FAN Y H, YAN D Q, et al. A novel removing irrelevant gene classification algorithm for tumor samples[J]. Journal

- of Liaoning Normal University(Natural Science Edition), 2015, 38(1): 41-46.
- [2] KOHANE I S, KHO A T, BUTTE A J. Micro arrays for an integrative genomics[M]. Cambridge: MIT Press, 2003.
 - [3] TOLOSI L, LENGAUER T. Classification with correlated features: unreliability of feature ranking and solutions[J]. Bioinformatics, 2011, 27(14): 1986-1994.
 - [4] 陈骥思, 余艳梅, 殷宇, 等. 自适应快速 FCM 彩色图像分割研究[J]. 计算机工程与应用, 2010, 46(7): 178-180.
CHEN J S, YU Y M, YIN Y, et al. Research on adaptive fast FCM color image segmentation[J]. Computer Engineering and Application, 2010, 46(7): 178-180.
 - [5] 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): 1226-1238.
 - [6] KOHAVI R. A study of cross-validation and boots trap for accuracy estimation and model selection[C]//The 14th International Joint Conference on Artificial Intelligence, August 20-25, 1995, Montréal, Canada. Burlington: Morgan Kaufmann Publishers, 1995: 1137-1143.
 - [7] HUANG D, LAI J H, WANG C D. Combining multiple clustering via crowd agreement estimation and multi-granularity link analysis [J]. Neuro Computing, 2015(170): 240-250.
 - [8] ALOM T N, SUDIPTA R. Identification of WBC based on dynamic clustering using modified FCM algorithm with an approach to optimal result[C]//ACM 6th International Conference on Computer and Communication Technology(ICCT 2015), September 25-27, 2015, Allahabad, India. New York: ACM Press, 2015: 461-464.
 - [9] 潘晓花, 孙文杰, 韦志辉, 等. 脑MR图像互信息最大的凸优化分割模型[J]. 计算机辅助设计与图形学学报, 2012, 24(8): 1082-1089.
PAN X H, SUN W J, WEI Z H, et al. The convex optimization segmentation model of MR brain image with maximum mutual information[J]. Journal of Computer Aided Design, 2012, 24(8): 1082-1089.
 - [10] SHIPP M A, ROSS K N, TAMAYO P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning[J]. Nat. Med, 2002(9): 68-74.
 - [11] DYRSKJOT L, THYKJAER T, KRUIHOFFER M, et al. Identifying distinct classes of bladder carcinoma using micro arrays[J]. Nature Genetics, 2003(33): 90-96.
 - [12] TROYANSKAYA O, CANTOR M, SHERLOCK G, et al. Missing value estimation methods for dna microarrays[J]. Bioinformatics, 2001, 17(6): 520-525.
 - [13] SINGH D, GFEBBO P, ROSS K, et al. Gene expression correlates of clinical prostate cancer behavior[J]. Cancer Cell, 2002, 1(2): 203-209.
 - [14] VAN'T VEER L J, DAI H, VANDEVIJVER M J, et al. Gene expression profiling predicts clinical outcome of breast cancer[J]. Nature, 2002, 415(6871): 530-536.
 - [15] POMEROY S L, TAMAYO P, GAASENBEEK M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression[J]. Nature, 2002, 415(6870): 436-442.
 - [16] GORDON G, JENSEN R, HSIAO L, et al. Translation of micro array data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma[J]. Cancer Res, 2002(62): 4963-4967.
 - [17] GUYON I, BENHUR A, GUNN S, et al. Result analysis of the nips 2003 feature selection challenge[C]//International Conference on Neural Information Processing Systems, November 22-25, 2004, Vancouver, British Columbia, Canada. New York: ACM Press, 2004: 545-552.
 - [18] 李鸿健, 代宇, 刘锐, 等. 云数据中心高能效的虚拟机迁移整合算法研究[J]. 电信科学, 2015, 31(1): 71-77.
LI H J, DAI Y, LIU R, et al. Energy-efficient virtual machine migration and consolidation algorithm in cloud data center[J]. Telecommunications Science, 2015, 31(1): 71-77.
 - [19] 张建敏, 谢伟良, 杨峰义, 等. 移动边缘计算技术及其本地分流方案[J]. 电信科学, 2016, 32(7): 132-139.
ZHANG J M, XIE W L, YANG F Y, et al. Mobile edge computing and application in traffic offloading[J]. Telecommunications Science, 2016, 32(7): 132-139.
 - [20] 李丹, 刘方明, 郭得科, 等. 软件定义的云数据中心网络基础理论与关键技术[J]. 电信科学, 2014, 30(6): 48-59.
LI D, LIU F M, GUO D K, et al. Fundamental theory and key technology of software defined cloud data center network[J]. Telecommunications Science, 2014, 30(6): 48-59.
 - [21] KALOUSIS A, PRADOS J, SANCHEZ J C, et al. Distilling classification models from cross validation runs: an application to mass spectrometry[C]//The 2004 IEEE International Conference on Tools with Artificial Intelligence, November 15-17, 2004, Boca Raton, Florida, USA. New Jersey: IEEE Press, 2004: 113-119.

[作者简介]



许小媛 (1980-), 女, 江苏开放大学信息与机电工程学院副教授, 主要研究方向为计算机软件应用、模式识别和算法研究。



黄黎 (1982-), 女, 南京航空航天大学计算机科学与技术学院博士生, 江苏开放大学信息与机电工程学院讲师, 主要研究方向为数据挖掘。