Advances in Applied Mathematics 应用数学进展, 2023, 12(1), 367-372
Published Online January 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.121040

Hans汉斯

半监督学习下的特征筛选问题研究

叶雨薇

南京信息工程大学，江苏 南京

收稿日期：2022年12月28日；录用日期：2023年1月21日；发布日期：2023年1月31日

摘　要

相比于传统的监督学习算法，半监督学习下的特征筛选算法可以利用更多已知信息提高模型计算性能。本文利用样本特征的分位数推测总体特征的分布情况，基于无模型假设下给出相对稳健的半监督特征筛选结果，模拟发现该算法在标记样本量相对较少且各类样本量不均衡的情况下适用。实例借助TCGA中的肺腺癌(LUAD)和肺鳞癌(LUSC)数据集验证算法的有效性。

关键词

特征筛选，半监督学习，分位数

Research on Feature Selection under Semi-Supervised Learning

Yuwei Ye

Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Dec. 28th, 2022; accepted: Jan. 21st, 2023; published: Jan. 31st, 2023

Abstract

Compared with the traditional supervised learning algorithm, the feature screening algorithm under semi-supervised learning can use more known information to improve the computational performance of the model. In this paper, the quantile of sample features is used to predict the distribution of overall features, and a relatively robust semi-supervised feature screening result is given based on the model-free hypothesis. Simulation results show that this algorithm is applicable when the number of labeled samples is relatively small and the number of various samples is unbalanced. The effectiveness of the algorithm is verified by the data sets of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) in TCGA.

文章引用: 叶雨薇. 半监督学习下的特征筛选问题研究[J]. 应用数学进展, 2023, 12(1): 367-372.
DOI: 10.12677/aam.2023.121040

Keywords

Feature Screening, Semi-Supervised Learning, Quantile

1. 引言

随着“互联网+”时代的到来，人类活动所产生的数据正以指数级爆炸式增长，利用数据信息辅助研究人员实现更为可靠的决策判断成为人民生活的迫切需求。特征筛选是数据分析和数据挖掘中的重要环节，其意义不仅在于显著减少运算时间、提高模型效率、降低过拟合风险、使回归或判别预测模型更为适用等，而且在很多情况下筛选出的特征具有一定的可解释性和潜在研究价值。

当前主流特征筛选方法为边际筛选方法，即仅考虑单个特征对响应变量的影响，不考虑特征之间的关系和对响应变量的影响。该方向算法整体上的优势在于运算速度快，处理超高维特征时相比于传统方法在计算的复杂性、准确性以及稳定性上效果更好。最早是由 Fan 和 Lv [1]提出利用皮尔逊相关系数搭建 SIS (Sure Independence Screening)算法，比较常用的有 Li 等[2]利用距离相关系数提出的 DC (Distance Correlation)-SIS 算法、Cui 等[3]利用条件分布函数和无条件分布函数之间的加权差构造的 MV (Market Value)算法以及 Mai 提出[4]利用 Kolmogorov-Smirnov (KS)检验进行的特征筛选思想并拓展到多分类情况下得到融合 Kolmogorov 过滤器(Fused Kolmogorov Filter, FKF)算法[5]。

仅用训练集中的特征和标签数据而忽略测试集中的特征数据是浪费的。这一问题在测试集中样本量远大于训练集中样本量时尤为突出。于是半监督学习应运而生，其常用思路可分为在无监督学习的基础上加入类别或成对约束和在监督学习的基础上加入特征之间的结构信息，目前已应用于图像处理和模式识别、音频语料分析、识别相关特征辅助医学诊断以及处理结构活性定量关系结构活性定量关系(Quantitative Structure Activity Relationship, QSAR)数据集辅助药物设计。He 等[6]率先提出利用局部和整体特征结构的一致性建立了无监督特征选择拉普拉斯计分(Laplacian Score, LS)算法为后续研究打下了基础，Zhao 等[7]将部分标签信息加入直接拓展至半监督领域得到 LSDF (Locality Sensitive Discriminant Analysis)算法，Cheng 等[8]则是在其基础上加入最小二乘和分类信息增益度(Classification Information Gain degree, CIG)构建 GSFS (Graph-based Semi-supervised Feature Selection)算法，Sheikhpour 等[9]针对回归问题基于 Laplacian 散点矩阵和正则化 L_2 范数框架上提出 SSCS (Semi-supervised Constraint Score)。

本文将 Song 等[10]在监督学习中的稳健加权分位数思想运用到半监督学习中，实现超高维数据下的半监督特征筛选问题，相比上述监督学习中的特征筛选算法具有更好的结果。

2. 半监督问题下的特征筛选算法

定义总体 (X,Y)，其中离散型响应变量 Y 有 R_n 个类别 $\{y_1,\cdots,y_{R_n}\}$，连续型特征变量 $X=(X_1,\cdots,X_p)^{\mathrm{T}}$。从总体中抽取 n 个有类标签的样本组成的样本集 $L=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，剩下的 $N-n(N>n)$ 个无类标签样本组成样本集 $U=\{x_{n+1},\cdots,x_N\}$，合并后得到总样本集 $T=L\cup U$。定义重要特征所在集合

$$D=\left\{j:\text{存在}\,y=y_r,\,X_j\,\text{与}\,F\left(y\mid X_j\right)\text{相关}\right\},$$

与之对应的不重要特征集合为 $I=\{1,2,\cdots,p\}\setminus D$。

定义第 j 个特征 X_j 下的 $\alpha\left(0<\alpha<1\right)$ 分位数为 $Q_\alpha\left(X_j\right)\triangleq\inf\left\{t:P\left(X_j\le t\right)\ge\alpha\right\}$，$Y=y_r$ 下 X_j 的 α 的分位数为 $Q_\alpha\left(X_j\mid Y=y_r\right)\triangleq\inf\left\{t:P\left(X_j\le t\mid Y=y_r\right)\ge\alpha\right\}$，$r=1,\cdots,R_n$，Song 等[10]在监督学习中综合得到加权分位指标 $w_j^\alpha\triangleq\sum_{r=1}^{R_n}p_r\left(Q_\alpha\left(X_j\mid Y=y_r\right)-Q_\alpha\left(X_j\right)\right)^2$，其中 $p_r=P\left(Y=y_r\right)$。又考虑到不能保证对任意 α，$w_j^\alpha=0$ 有 Y 和 X_j 是独立的，给出稳健加权分位数

$$w_j=\int_0^1\sum_{r=1}^{R_n}p_r\left(Q_\alpha\left(X_j\mid Y=y_r\right)-Q_\alpha\left(X_j\right)\right)^2\mathrm{d}\alpha,j=1,\cdots,p.\tag{1}$$

本文将公式(1)从监督学习背景推广至半监督学习，多个分位点取代连续型积分。取 $M=10$ 对区间 $[0,1]$ 进行切片，此时可给出利用数据估计的半监督稳健加权分位数(Semi-supervised Robust Composite Weighted Quantile, SRCWQ)

$$\hat{w}_j^{(Q)}=\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{R_n}\hat{p}_r\left(\hat{Q}_{\alpha_k}^L\left(X_j\mid Y=y_r\right)-\hat{Q}_{\alpha_k}^T\left(X_j\right)\right)^2,\tag{2}$$

其中，$\alpha_k=\left(0.5+k-1\right)/M,k=1,\cdots,M$，$\hat{p}_r=\frac{1}{n}\sum_{i=1}^{n}I\{Y=y_r\}$。总样本集 T 下经验分布函数为 $\hat{F}_{N,X_j}\left(x\right)\triangleq\frac{1}{N}\sum_{i=1}^{N}I\left\{X_{ij}\le x\right\}$，则对应分位数为 $\hat{Q}_{\alpha_k}^T\left(X_j\right)\triangleq\hat{F}_{N,X_j}^{-1}\left(\alpha_k\right)$；已标记样本集 L 下经验条件分布函数为 $\hat{F}_{n,X_j\mid Y=y_r}\left(x\right)\triangleq\frac{\sum_{i=1}^{n}I\left\{X_{ij}\le x,Y=y_r\right\}}{\sum_{i=1}^{n}I\left\{Y=y_r\right\}}$，对应分位数 $\hat{Q}_{\alpha_k}^L\left(X_j\mid Y=y_r\right)\overset{\Delta}{=}\hat{F}_{n,X_j\mid Y=y_r}^{-1}\left(\alpha_k\right)$。

本文根据按照从大到小的顺序对 p 个特征的 SRCWQ 指标值进行排序，取前 $d=\left[n/\log\left(n\right)\right]$ 个组成重要特征集合 D 即为特征筛选所得结果。

3. 数值模拟

本文采用蒙特卡洛算法进行模拟，目标是对比不同情况下不同算法的特征筛选方法优劣。

取响应变量 $Y=0$ 或 1，其中 $Y=y_r$ 下有重要特征 τ_{y_r}，$X_i=\mu_{\tau_{y_r}}+\varepsilon_i$，$\mu_{\tau_{y_r}}$ 即表示除第 τ_{y_r} 个为 1 其余均为 0 的 p 维向量，$\varepsilon\sim N\left(0,\Sigma\right),\Sigma=I_p$。设置默认参数：随机数种子 100、已标记样本量个数 $n=50$、总样本量个数 $N=300$、重抽样次数 $B=200$，特征维数 $p=2000$ 以及均衡状态 $P\left(Y=0\right)=P\left(Y=1\right)$。

将本文提出的半监督学习算法 SRCWQ 与传统监督学习算法 DC [2]、MV [3]以及 FKF [5]进行对比，利用以上参数构造模型 1，从均衡状态、异常值以及协方差三个角度修改参数得到模型 2~4。其中，模型 2 修改模型 1 中均衡状态为 $P\left(Y=0\right):P\left(Y=1\right)=4:1$，模型 3 为模型 1 中增加 4 个异常点，模型 4 修改模型 1 中协方差 $\Sigma=\left(\sigma_{ab}\right)_{p\times p},\sigma_{ab}=0.3^{|a-b|},1\le a,b\le p$。

构造模型的评价指标 $MMS\left(a\right)$、RSD、P_k 以及 P_a。在 $B=200$ 次循环中，每一次选中全部重要特征所需要选择的特征数量为 $MMS_b\left(a\right),b=1,\cdots,B$，$RSD$ 为 $MMS\left(a\right)$ 上下四分位数之差除以 1.34，表示特征筛选结果的稳定性；P_k 表示在前 $d=\left[n/\log\left(n\right)\right]$ 个所选特征中包含第 k 个重要特征的比例，P_a 表示在前 $d=\left[n/\log\left(n\right)\right]$ 个所选特征中包含全部重要特征的比例。根据上述定义，$MMS\left(a\right)$ 和 RSD 的数值越小越好，而 P_k 和 P_a 则是越大越好。表 1 利用上述指标给出不同算法下不同模型的特征筛选结果情况，可以看到在样本不均衡分布和存在异常点时效果较好。

Table 1. Summary of feature screening results under different algorithms
表 1. 不同算法下特征筛选结果情况汇总

	$MMS(a)$				RSD	P_1	P_2	P_a
	25%	50%	75%	100%				
模型 1：默认参数								
SRCWQ	2	2	2	86	0.00	0.995	0.99	0.985
DC	2	2	2	64	0.00	0.97	0.98	0.95
MV	2	2	3	84	0.75	0.975	0.99	0.965
FKF	2	2	4	147	1.49	0.935	0.96	0.9
模型 2：不均衡分布								
SRCWQ	2	2	7	235	3.73	0.93	0.885	0.825
DC	2	4	13	436	8.21	0.88	0.81	0.715
MV	2	5	18.25	323	12.13	0.845	0.8	0.665
FKF	3	11	29.5	498	19.78	0.79	0.675	0.51
模型 3：增加异常点								
SRCWQ	2	2	2	32	0.00	0.99	0.995	0.985
DC	2	3	7	900	3.73	0.825	0.975	0.8
MV	2	2	4.25	408	1.68	0.9	0.965	0.865
FKF	2	3	11	329	6.72	0.815	0.935	0.76
模型 4：改变协方差								
SRCWQ	2	2	2	49	0.00	0.99	0.985	0.975
DC	2	2	2	133	0.00	0.985	0.985	0.97
MV	2	2	2	192	0.00	0.985	0.985	0.97
FKF	2	2	3	569	0.75	0.97	0.945	0.92

4. 实例分析

本文利用公开数据集 TCGA 中肺腺癌(LUAD)和肺鳞癌(LUSC)的基因表达 RNA 序列数据集构造半监督数据集。原始数据来源于网址 https://gdc.xenahubs.net，仅提取尾号为“01A”(表示患病)和“11A”(表示正常)的样本。考虑到基因探针信息一致，两数据集可以进行合并，总特征个数 $p=60488$，其中 60483 个为基因编号，5 个样本基因评价指标分别为 alignment_not_unique (比对位置不唯一)、ambiguous (比对区域与多个基因都发生重叠)、no_feature (比对区域与任何基因都没有重叠)、not_aligned (无法比对上)以及 too_low_aQual (比对质量低于设定阈值，默认是 10)。从数据集中的样本名称提取活体组织检测结果，

得到样本分类 LUAD (对应类别 1)、LUSC (对应类别 2)以及 Normal (正常，对应类别 0)共 3 种情况，每类样本量分别为 510、496 以及 107。于是可得总样本集 T 中样本量 $N=1113$，按照类样本比例抽取 5% 的样本构成已标记样本集 L，筛选特征个数 $d=13$，重复抽样实验 200 次降低偶然性。

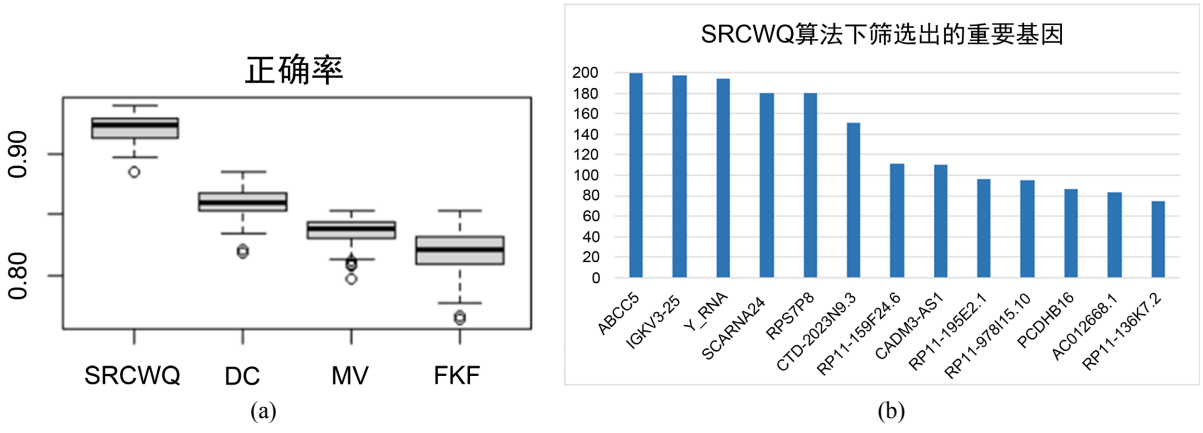
Table 2. Confusion matrix under three categories
表 2. 三分类下的混淆矩阵

预测 \ 真实	0	1	2
0	a_1	a_4	a_7
1	a_2	a_5	a_8
2	a_3	a_6	a_9

构造三分类混淆矩阵如表 2 所示，给出模型评价指标定义如下：

- 正确率：$(a_1+a_5+a_9)/(a_1+a_2+a_3+a_4+a_5+a_6+a_7+a_8+a_9)$，表示模型分类性能，得分越高越好。
- LUAD 误诊率：$1-a_5/(a_4+a_5+a_6)$，表示样本患 LUAD 但是诊断错误的概率，得分越低越好。
- LUSC 误诊率：$1-a_9/(a_7+a_8+a_9)$，表示样本患 LUSC 但是诊断错误的概率，得分越低越好。
- LUAD 和 LUSC 误诊率：$1-(a_5+a_9)/(a_5+a_6+a_8+a_9)$，表示样本患 LUAD 但被诊断为 LUSC 或样本患 LUSC 但被诊断为 LUAD 的概率，得分越低越好。
- 是否患病误诊率：$(a_2+a_3+a_4+a_7)/(a_1+a_2+a_3+a_4+a_5+a_6+a_7+a_8+a_9)$，表示样本患病但未被诊断出患病或样本不患病但被诊断出患病的概率，得分越低越好。

给出 200 次抽样实验中，每次选中的前 13 个特征结果。对特征频次进行从大到小排序后，选择前 13 个特征进行 200 次随机森林分类，给出每次分类下正确率结果，每个模型情况如图 1(a)所示。其中可以看到 SRCWQ 在单次分类准确率和稳定性上优于其它算法。该算法下筛选出的重要基因如图 1(b)所示，认为若该基因出现频次越高，则致病可能性越大。



(a) (b)

Figure 1. Schematic diagram of model evaluation under different algorithms
图 1. 不同算法下的模型评价结果示意图

考虑到现实生活中数据的不完整性，对 200 次抽样实验中的每一次特征筛选结果做一次随机森林分类，给出每次分类下的误诊率如图 2 所示，从中可以看到 SRCWQ 相比于其它算法有更少的可能会出现误诊情况，尤其在是否患病的情况下较为突出。

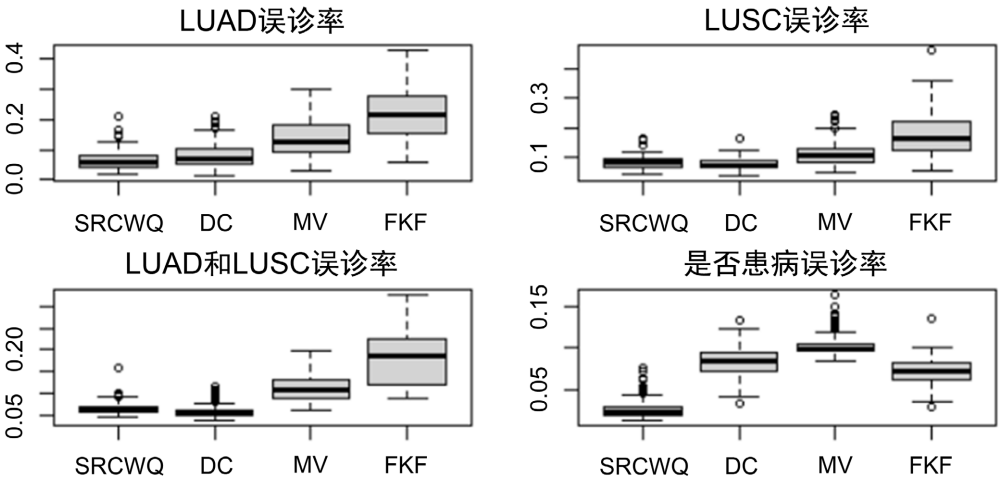


Figure 2. Misdiagnosis rate under different algorithms
图 2. 不同算法下的误诊率

5. 总结

综上，本文在 Song 等[10]基础上进行拓展，利用样本特征的分位数分布，提出了一种半监督学习下的特征筛选方法 SRCWQ。经数值模拟验证，该算法在已标记样本量显著小于总样本量下，针对各类样本量不均衡或存在异常点的情况较为适用。经实例验证，该算法可借助大量无标记样本，对罕见病诊疗有一定的辅助判断作用。

参考文献

[1] Fan, J. and Lv, J. (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **70**, 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

[2] Li, R., Zhong, W. and Zhu, L. (2012) Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, **107**, 1129-1139. https://doi.org/10.1080/01621459.2012.695654

[3] Cui, H., Li, R. and Zhong, W. (2015) Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *Journal of the American Statistical Association*, **110**, 630-641. https://doi.org/10.1080/01621459.2014.920256

[4] Mai, Q. and Zou, H. (2013). The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification. *Biometrika*, **100**, 229-234. https://doi.org/10.1093/biomet/ass062

[5] Mai, Q. and Zou, H. (2015) The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method. *The Annalsof Statistics*, **43**, 1471-1497. https://doi.org/10.1214/14-AOS1303

[6] He, X., Cai, D. and Niyogi, P. (2005) Laplacian Score for Feature Selection. *Advances in Neural Information Processing Systems*, **18**, 507-514.

[7] Zhao, J., Lu, K. and He, X. (2008) Locality Sensitive Semi-Supervised Feature Selection. *Neurocomputing*, **71**, 1842-1849. https://doi.org/10.1016/j.neucom.2007.06.014

[8] Cheng, H., Deng, W., Fu, C., Wang, Y. and Qin, Z. (2011) Graph-Based Semi-Supervised Feature Selection with Application to Automatic Spam Image Identification. In: Yu, Y., Yu, Z. and Zhao, J., Eds., *Computer Science for Environmental Engineering and EcoInformatics*. *CSEEE* 2011. *Communications in Computer and Information Science*. Springer, Berlin, Heidelberg, 259-264. https://doi.org/10.1007/978-3-642-22691-5_45

[9] Sheikhpour, R., Sarram, M.A. and Sheikhpour, E. (2018) Semi-Supervised Sparse Feature Selection via Graph Laplacian Based Scatter Matrix for Regression Problems. *Information Sciences*, **468**, 14-28. https://doi.org/10.1007/978-3-642-22691-5_45

[10] Song, F., Lai, P. and Shen, B. (2020). Robust Composite Weighted Quantile Screening for Ultrahigh Dimensional Discriminant Analysis. *Metrika*, **83**, 799-820. https://doi.org/10.1007/s00184-019-00758-x