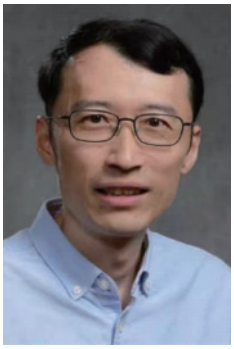

【专家简介】秦国友,博士,美国北卡罗莱纳大学教堂山分校博士后。现为复旦大学教授、博士生导师,公共卫生学院生物统计学教研室主任。主要从事生物统计学方法学和应用研究,包括真实世界研究和因果推断,针对复杂数据、复杂统计模型的统计方法创新,以及生物统计学方法在医学和公共卫生领域的应用。纵向数据分析相关研究工作获得2014年教育部高等学校科学研究优秀成果二等奖。在*BMJ*,*PLoS Medicine*,*JAMA Network Open*,*Biometrics*,*Biostatistics*和*Statistics in Medicine*等医学和生物统计学权威期刊上发表论文100余篇。主要学术兼职:中华预防医学会生物统计学分会第一届青年委员会主任委员,中国卫生信息学会健康统计专业委员会和中华预防医学会生物统计分会常务委员,中国现场统计研究会多元分析应用专业委员会与全国工业统计学教学研究会健康医疗大数据学会常务理事,以及《中国卫生统计》《中国预防医学》杂志编委。

# 迁移学习简介及其在医学研究领域中的应用

潘璐璐　余勇夫　秦国友△

(复旦大学公共卫生学院生物统计学教研室　上海　200032)

【摘要】 本文介绍了一种基于回归模型的迁移学习,并通过实例数据展示了其在医学领域的应用。实例基于美国健康和营养调查2013—2014年的数据,研究了睡眠时间和抑郁程度及抑郁症之间的关联,并使用人口特征和生活方式作为预测变量,预测不同种族群体的抑郁程度和抑郁症。相比只基于目标种族群体构建的模型,迁移学习可以提高目标种族中睡眠时间效应的估计精度,以及抑郁程度和抑郁症的预测准确性。实例结果表明,在目标数据稀缺和数据源之间存在异质性的情况下,迁移学习能够有效整合外源数据,显著提升目标模型的估计能力和预测能力。

【关键词】 迁移学习;　估计;　预测

【中图分类号】 R311　【文献标志码】 A　**doi**:10.3969/j.issn.1672-8467.2024.06.020

# Introduction and application of transfer learning in medical research

PAN Lu-lu, YU Yong-fu, QIN Guo-you△

(*Department of Biostatistics*, *School of Public Health*, *Fudan University*, *Shanghai* 200032, *China*)

【**Abstract**】 This paper introduces a transfer learning approach based on regression models and demonstrates its application in the medical field through an example. Using data from the 2013–2014 U.S. National Health and Nutrition Examination Survey, the study investigates the association of sleep duration with depression levels and depressive disorder. It employs demographic characteristics and lifestyle factors as predictor variables to predict depression levels and depressive disorder across different racial groups. Compared to models built solely on target racial groups, transfer learning enhances the accuracy of estimating the effect of sleep duration in the target group and improves the prediction accuracy for depression levels and depressive disorder. The results illustrate that transfer learning effectively integrates

国家自然科学基金(82173612);上海市市级科技重大专项(ZD2021CY001)

△Corresponding author　E-mail: gyqin@fudan.edu.cn

网络首发时间:2024-11-21 14:25:51　网络首发地址:https://link.cnki.net/urlid/31.1885.R.20241121.0842.002

source data to significantly improve estimation and prediction capabilities of target models, especially in situations with limited target data and heterogeneous data sources.

【Key words】 transfer learning; estimation; prediction

*** This work was supported by the National Natural Science Foundation of China (82173612) and Shanghai Municipal Science and Technology Major Project (ZD2021CY001).**

随着临床研究和公共健康数据的不断积累，医学研究获得了大量关于人群健康状况、生活方式及社会经济背景的信息。这些数据为疾病的早期诊断、个性化治疗和预后评估提供了新的研究视角和研究机会。然而，生物医学和临床医学研究中常因伦理限制和高昂的研究成本，难以获取足够的生物学或临床样本，这限制了研究结果的精确性[1-2]。此外，来自不同医院或研究机构的数据往往在实验设置、患者特征等方面存在显著差异，导致数据源之间异质性较大，直接合并数据通常难以充分反映目标数据的特征[3-5]。

为了应对这一挑战，迁移学习近年来在医学领域得到了广泛关注[6]。既往研究已提出了基于回归模型的迁移学习框架，核心在于将外源数据中的有用信息迁移至目标数据，从而提高目标模型中回归系数估计的准确性。迁移学习不仅能够更精准地估计风险因素的效应，还能提升对结局的预测效果。如Bastani在2021年提出了一种基于线性回归模型的灵活两步迁移学习框架，通过结合一个信息丰富的外源数据，显著提升了目标数据的预测能力[7]。Li等[8-9]和Tian等[10]进一步提出了能考虑多个外源数据的迁移学习框架，并将该框架推广至广义线性模型，显著提高了目标模型的估计能力和预测能力。Pan等[11]在前述Li等的研究基础上提出了稳健的迁移学习，旨在克服现实世界中常见的异常值问题。Zhu等[12]进一步将迁移学习框架推广至分位数回归模型，以更好地捕捉因变量在不同分位数下的异质性效应，从而提升目标群体中感兴趣变量与结局之间关联估计的精确性和结局预测的准确性。这些研究体现了迁移学习在效应估计和结局预测上均具备显著优势。因此，迁移学习在实际应用中的重要性日益突出。

本文介绍了一种基于回归模型的迁移学习框架，通过估计睡眠时间与抑郁之间的关联，并预测不同种族群体的抑郁程度和抑郁症的实例，展示了迁移学习在医学数据分析中的优势，为后续研究和实践提供参考。

**迁移学习框架** 本文基于Li等[8]的研究，介绍一种基于回归模型的迁移学习框架。以线性回归模型为例阐述该框架的核心思想。当因变量为连续型变量时，目标模型为线性回归模型，表示为：$y_i^{(0)}=\left(x_i^{(0)}\right)^T\beta+\varepsilon_i^{(0)}, i=1,\cdots,n_0$；其中$\beta\in\mathbb{R}^p$是回归系数，$\boldsymbol{\varepsilon}_i^{(0)}$是服从正态分布的误差项。

假定有$K$个外源数据，第$k$个外源数据对应的回归模型可以表示为$y_i^{(k)}=\left(x_i^{(k)}\right)^T\omega^{(k)}+\varepsilon_i^{(k)}, i=1,\cdots,n_k, k=1,\cdots,K$，其中回归系数$\omega^{(k)}$与目标模型的回归系数$\beta$不同。假定向量$\delta^{(k)}=\beta-\omega^{(k)}$表示$\beta$与$\omega^{(k)}$之间的差异向量，当差异$\delta^{(k)}$足够小时，则认为该数据集是可迁移的，可以通过迁移学习利用该数据的信息提升目标数据的学习效果。

基于回归模型的两步迁移学习框架包含以下两个步骤：第一步，使用目标数据$\left(X^{(0)},Y^{(0)}\right)$和所有可迁移的外源数据$\{X^{(k)},Y^{(k)}\}_{k=1}^K$计算回归系数向量$\hat{\omega}$，$\hat{\omega}=\underset{\omega\in\mathbb{R}^p}{\arg\min}\left\{L\left(\omega;X^{(k)},Y^{(k)}\right)\right\}, k=0,\cdots,K$。其中，$L(\omega;X^{(k)},Y^{(k)})$为线性回归模型的损失函数，表示为$\frac{1}{n}\sum_{i=1}^{n}(y_i^{(k)}-x_i^{(k)T}\omega)^2$，$n=\sum_{k=0}^{K}n_k$是目标数据和所有可迁移的外源数据的样本总数。$\hat{\omega}$通常是一个有偏的向量，因此需要对$\hat{\omega}$的偏倚进行纠正。第二步，使用目标数据$\left(X^{(0)},Y^{(0)}\right)$纠正$\hat{\omega}$的偏倚。定义$\hat{\omega}$和目标数据真实回归系数$\beta$之间的差异向量为$\delta$，$\hat{\delta}=\underset{\delta\in\mathbb{R}^p}{\arg\min}\left\{L(\delta+\hat{\omega};X^{(0)},Y^{(0)})+\lambda_\delta\|\delta\|_1\right\}$，其中，$\lambda_\delta=\sqrt{logp/n_0}$，$\|\delta\|_1$是$\delta$的L-1范数（即参数的绝对值之和）是常用的正则化方法，可以使部分参数收缩到零，这在特征选择和提升模型解释性方面具有优势。最终的回归系数估计为$\hat{\beta}=\hat{\omega}+\hat{\delta}$。

此外，该迁移学习框架也可以扩展至处理分类问题的logistic回归模型。此时目标模型可写成：$logit\{P(y_i^{(0)}=1|x_i^{(0)})\}=\left(x_i^{(0)}\right)^{\mathrm{T}}\beta, i=1,\cdots,n_0$。第$k$个外

源数据对应的回归模型可以表示为 $logit\{P(y_i^{(k)}=1|x_i^{(k)})\}=\left(x_i^{(k)}\right)^{\mathrm{T}}\omega^{(k)}, i=1,\cdots,n_k, k=1,\cdots,K$。logistic回归模型的损失函数 $L(\omega;X^{(k)},Y^{(k)})$ 可表示为 $\frac{1}{n}\sum_{i=1}^{n}(log\{1+exp(x_i^{(k)\mathrm{T}}\omega)\}-y_i^{(k)}x_i^{(k)\mathrm{T}}\omega)$。这两种模型的介绍体现了迁移学习在处理不同类型因变量时的适用性。迁移学习可以通过Tian等[10]编写的R包*glmtrans*实现,该包为基于线性回归和逻辑回归等广义线性模型的迁移学习框架提供了便捷的分析工具。

**实例分析** 抑郁症作为全球范围内常见且具有严重负担的心理健康疾病,对个体的生理、心理和社会功能会产生广泛的影响[13]。不同种族群体的抑郁症表现和病理机制可能存在显著差异。种族、文化、社会经济状况以及健康行为等因素都会影响个体的抑郁症风险[14]。例如,某一特定种族群体会受其独特的社会经济背景或文化习惯的影响,而这些因素在现有的全人群模型中可能未被充分考虑或被低估。因此,基于全人群构建的模型在不同种族群体中的适用性和效果可能会有差异。为特定种族群体构建独立的抑郁风险模型,不仅有助于更准确地估计该群体的独特风险因素和疾病模式,提供更加精准的预防和干预措施,还能显著提升结局预测的准确性和可靠性,从而更有效地识别抑郁症高风险个体。然而,现实研究中单一种族群体的数据样本通常较为有限,往往难以构建足够精确的风险模型和预测模型。因此,本研究应用迁移学习,将来自其他种族群体的外源数据迁移至目标种族群体,以提高目标群体中风险因素与抑郁得分和抑郁症关联估计的精确性以及抑郁得分和抑郁症预测的准确性。

本研究的数据来自于2013年至2014年美国健康和营养调查(National Health and Nutrition Examination Survey, NHANES)公开数据库(https://www.cdc.gov/nchs/nhanes/index.htm)[15]。NHANES是一项基于人群的横断面调查,旨在收集美国成人和儿童的健康和营养状况的信息,采用患者健康问卷(Patient Health Questionnaire-9, PHQ-9)来评估个体的抑郁程度。该问卷包括9个问题,调查过去2周内抑郁症状的频率,得到范围为0~27的抑郁得分作为连续型的因变量,根据总分≥10作为临床抑郁症的判定标准,将抑郁得分划分为二分类的因变量[16]。感兴趣的自变量为睡眠时间,由调查对象自我回忆并报告,定义为工作日晚上平均睡眠时间,<6 h和>9 h被定义为短睡眠时间和长睡眠时间。协变量包括年龄、性别、教育程度、家庭收入贫困比、婚姻状况、体育活动及酒精消费状况。研究对象的年龄范围为20~79岁,排除缺失睡眠时间、抑郁得分和重要协变量的个体,使用完整数据集进行分析。本研究将NHANES中的5个种族群体(墨西哥裔美国人、其他西班牙裔、非西班牙裔白人、非西班牙裔黑人、其他种族)分别作为目标数据群体,其他4个种族作为外源数据进行迁移学习。研究有两个目的:一是估计睡眠时间与抑郁程度和抑郁症之间的关联效应;二是使用感兴趣的自变量和协变量作为预测变量,以构建抑郁程度和抑郁症的预测模型。研究通过比较只基于目标数据构建的模型与迁移学习的模型,评估迁移学习在提高模型估计精度和增强预测能力方面的效果。

表1和表2展示了在5个种族中,睡眠与抑郁程度和抑郁症之间关联效应的估计值和标准误差。与仅基于目标数据构建的多因素回归模型相比,迁移学习可以明显降低回归系数估计的标准误差,提高效应估计的精确性。图1和图2展示了在5个种族中抑郁程度的相对预测误差和抑郁症的受试者工作特征曲线下面积(area under the ROC curve, AUC)。结果表明,与仅基于目标数据的预测模型相比,迁移学习在大多数情况下都能明显提高抑郁程度和抑郁症的预测效果(平均降低14%的预测误差和提高12%的AUC)。这些结果突显了整合外源数据的有用信息在提升目标数据的估计和预测方面的优势。

**总结** 本文介绍了一种基于回归模型的迁移学习框架。该方法的关键优势在于,当目标数据量有限时,可以迁移外源数据中的有用信息,从而有效提升目标数据的估计和预测能力。通过估计睡眠时间与抑郁之间的关联,并预测不同种族群体的抑郁程度和抑郁症的实例,展示了迁移学习在医学领域的应用潜力。

现实中不同种族群体在社会背景、文化习惯、健康状况等方面存在显著差异,这些因素可能导致传统模型在某些种族群体中的估计精度和预测效果较低。通过迁移学习,可以借用其他种族群体的相关数据作为外源信息,帮助模型克服目标数据中

表1 NHANES中5个种族睡眠与抑郁程度关联的回归系数（标准误）

Tab 1 Regression coefficients (standard errors) for the association between sleep and depression severity across five races in NHANES

| Race | Target data | | Transfer with sources | |
|---|---|---|---|---|
| | Short sleep duration | Long sleep duration | Short sleep duration | Long sleep duration |
| Mexican American | 1.46 (0.39) | 0.39 (0.42) | 1.27 (0.30) | 0.24 (0.21) |
| Other Hispanic | 0.53 (0.68) | −0.85 (0.70) | 0.76 (0.46) | −0.12 (0.35) |
| Non-Hispanic white | 0.96 (0.20) | 0.30 (0.21) | 1.11 (0.19) | 0.29 (0.11) |
| Non-Hispanic black | 1.10 (0.34) | −0.66 (0.36) | 1.07 (0.30) | −0.07 (0.17) |
| Other races | 1.09 (0.35) | 0.48 (0.36) | 1.01 (0.34) | 0.40 (0.20) |

表2 NHANES中5个种族睡眠与抑郁症关联的回归系数（标准误）

Tab 2 Regression coefficients (standard errors) for the association between sleep and major depressive disorder across five races in NHANES

| Race | Target data | | Transfer with sources | |
|---|---|---|---|---|
| | Short sleep duration | Long sleep duration | Short sleep duration | Long sleep duration |
| Mexican American | 0.95 (0.49) | 0.86 (0.51) | 0.55 (0.32) | 0.27 (0.45) |
| Other Hispanic | 0.47 (0.43) | −0.71 (0.60) | 0.66 (0.24) | −0.12 (0.43) |
| Non-Hispanic white | 0.44 (0.23) | 0.17 (0.23) | 0.50 (0.16) | 0.03 (0.19) |
| Non-Hispanic black | 0.53 (0.31) | −0.22 (0.34) | 0.54 (0.16) | −0.08 (0.25) |
| Other races | 0.83 (0.66) | 0.82 (0.67) | 0.57 (0.40) | 0.12 (0.54) |

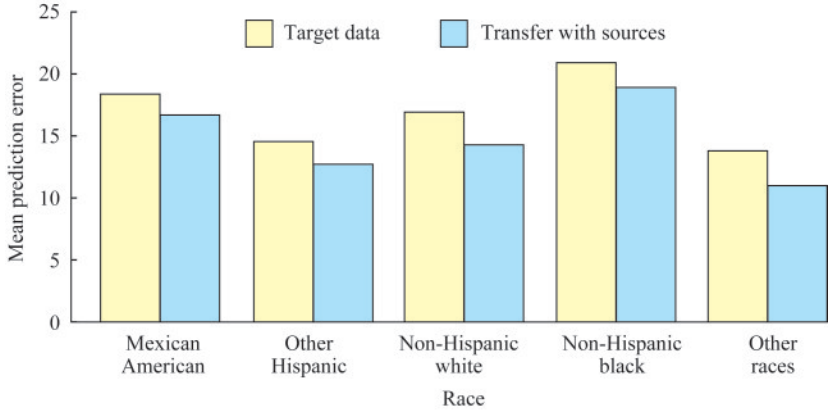


图1 NHANE中5个种族抑郁程度的平均预测误差

Fig 1 Mean prediction error of depression severity across five races in NHANES

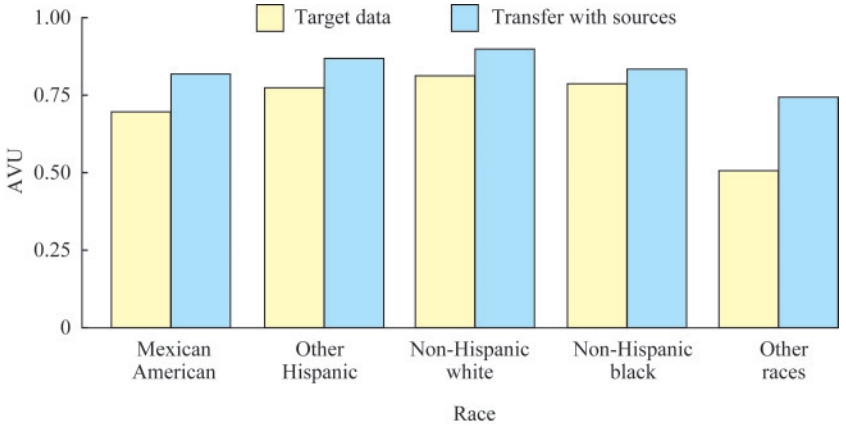


图2 NHANES中5个种族抑郁症的AUC

Fig 2 AUC of major depressive disorder across five races in NHANES

样本不足的问题,提升在特定种族群体中的效应估计精确性和预测能力。特别是在涉及到少数族裔群体时,迁移学习能够有效解决数据稀缺的问题,使得模型在这些群体中同样提供可靠的效应估计和准确的预测,这为实现更加公平和全面的健康管理提供了新的技术路径。需要指出的是,尽管迁移学习能够有效利用外源数据提高估计和预测能力,但也应注意到负迁移的潜在风险。负迁移发生在外源数据与目标数据之间存在显著差异时,可能导致模型性能下降[17]。因此,在应用迁移学习时,我们需要谨慎评估外源数据与目标数据的相似性,并采取适当的策略来减少负迁移的影响。尽管如此,迁移学习在提升特定群体中估计和预测能力方面的优势依然明显。

综上所述,迁移学习在医疗研究领域的数据整合方面具有巨大潜力,能够帮助研究者将外源数据中的有用信息应用到目标数据中,在有限的目标数据条件下获得更准确的分析结果,从而为个性化治疗和疾病预测提供有力支持。

**作者贡献声明** 潘璐璐 数据整理、结果分析和解释,论文撰写。余勇夫 课题构思和设计,论文修订。秦国友 研究设计和指导,资助获取,论文修订。

**利益冲突声明** 所有作者均声明不存在利益冲突。

## 参 考 文 献

[1] HAMID JS, HU P, ROSLIN NM, *et al*. Data integration in genetics and genomics: methods and challenges[J]. *Hum Genomics Proteomics*, 2009, 2009: 869093.

[2] GOMEZ-CABRERO D, ABUGESSAISA I, MAIER D, *et al*. Data integration in the era of omics: current and future challenges[J]. *BMC Syst Biol*, 2014, 8(Suppl 2): I1.

[3] ALYASS A, TURCOTTE M, MEYRE D. From big data analysis to personalized medicine for all: challenges and opportunities[J]. *BMC Med Genomics*, 2015, 8: 33.

[4] BANSAL S, SINDHI V, SINGLA BS. Exploration of deep learning and transfer learning techniques in bioinformatics[M]//Applying machine learning techniques to bioinformatics: few-shot and zero-shot methods. Hershey, Pennsylvania: IGI Global, 2024: 238-257.

[5] GANGWAL A, ANSARI A, AHMAD I, *et al*. Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review[J]. *Comput Biol Med*, 2024, 179: 108734.

[6] WEISS K, KHOSHGOFTAAR TM, WANG D. A survey of transfer learning[J]. *J Big Data*, 2016, 3(1): 1-40.

[7] BASTANI H. Predicting with proxies: transfer learning in high dimension[J]. *Management Sci*, 2021, 67(5): 2964-2984.

[8] LI S, CAI TT, LI H. Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality[J]. *J R Stat Soc Ser B Stat Methodol*, 2022, 84(1): 149-173.

[9] LI S, ZHANG L, CAI TT, *et al*. Estimation and inference for high-dimensional generalized linear models with knowledge transfer[J]. *J Am Stati Assoc*, 2024, 119(546): 1274-1285

[10] TIAN Y, FENG Y. Transfer learning under high-dimensional generalized linear models[J]. *J Am Stat Assoc*, 2023, 118(544): 2684-2697.

[11] PAN L, GAO Q, WEI K, *et al*. A robust transfer learning approach for high-dimensional linear regression to support integration of multi-source gene expression data[J]. *PLoS Comput Biol* [In Revision], 2024.

[12] ZHANG Y, ZHU Z. Transfer learning for high-dimensional quantile regression *via* convolution smoothing[J]. *Statistica Sinica*[Preprint No: SS-2022-0396].

[13] Smith K. Mental health: a world of depression[J]. *Nature*, 2014, 515(7526): 181.

[14] MONTALVO-LIENDO N, GROGAN-KAYLOR A, GRAHAM-BERMANN S. Ethnoracial variation in depression symptoms[J]. *Hisp Health Care Int*, 2016, 14(2): 81-88.

[15] DILLON CF, WEISMAN MH. US National Health and Nutrition Examination Survey Arthritis initiatives, methodologies and data[J]. *Rheum Dis Clin North Am*, 2018, 44(2): 215-265.

[16] LIU X, LIU X, WANG Y, *et al*. Association between depression and oxidative balance score: National Health and Nutrition Examination Survey (NHANES) 2005-2018[J]. *J Affect Disord*, 2023, 337: 57-65.

[17] WANG Z, DAI Z, PÓCZOS B, *et al*. Characterizing and avoiding negative transfer [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, 2019, 11285-11294.

(收稿日期:2024-11-12; 编辑:张秀峰)