鲁东大学学报(自然科学版)
Journal of Ludong University(Natural Science Edition) 2021,37(1):1—5

基于 Logistic 回归和 Noisy-or 模型的抑郁症风险预测研究

杨 斐,魏新江

(鲁东大学 数学与统计科学学院,山东 烟台 264039)

摘要: 本文为了实现抑郁症的早期识别和检测,将 Logistic 回归模型与 Noisy-or 模型相结合,提出一种用于抑郁症早期识别和检测的风险预测模型 LRANO。在该模型中,Logistic 回归用于计算不同风险因素对抑郁的概率贡献,Noisy-or 模型将各种模型参数整合,形成最终的抑郁症风险预测模型。此外,通过在爱尔兰老龄化纵向研究数据库(TILDA)中进行验证,该模型的 AUC 值为0.731 3,平均绝对误差为0.288 7,表明了模型的有效性。

关键词: 抑郁症; Logistic 回归; Noisy-or; 风险预测

中图分类号: O213; R749 **文献标志码:** A **文章编号:** 1673-8020(2021)01-0001-05

抑郁症是一种常见的精神疾病。近年来,人们经历的来自学习、生活和社会的压力越来越大,导致抑郁症的发病率逐年升高[1]。据世界卫生组织(WHO)统计,全世界有超过 3 亿人患有抑郁症。目前,抑郁症已成为世界第二大致残疾病,严重威胁人类的身心健康,被称为心理病理中的普通感冒[2—3]。

近年来,越来越多的人意识到抑郁症的危害及其早期发现的重要性。抑郁症具有治愈率低、易复发、死亡率和致残率高的特点。重度抑郁症患者的后果更为严重,如丧失行为能力、恶化各种身体指标、增加医疗费用和自杀风险等[4]。虽然抑郁症的治疗方法多种多样,但由于医疗资源的缺乏以及抑郁症早期发现较困难,其治疗效果较差[3,5]。因此,建立抑郁症的风险预测模型,实现抑郁症的早期发现是非常重要的。

目前,关于抑郁症的研究主要是探讨风险因素[6—8],对风险预测模型的研究相对较少。文献[9]通过对生活在韩国安山市的 299 名老年人进行私人访谈获得预测变量,建立了基于路径分析的老年抑郁症预测假设模型;文献[10]运用线性判别分析建立日本社区老年人的抑郁症预测模型,其中预测因素包括听力问题、食欲、经济、情绪问题和主观有用性。文献[9]和[10]基于横截面数据进行研究,预测效果较差。文献[11]通过对纵向数据的分析研究,建立了抑郁风险评估工具 DRAT-up,该工具可用于评估存在缺失值时的抑郁风险。事实上,风险评估工具 DRAT-up 中风险因素的影响是在一种过时且复杂的系统回顾和元分析的基础上得到的。因此,该工具中的参数值不能适应数据的变化,不能很好地达到预测效果。

本文为解决抑郁症的风险预测问题,建立一种既能应用于纵向数据,又能适应数据变化且提高预测精度的模型。Logistic 回归既能够找到对抑郁症具有显著影响的风险因素,又可以提供每个风险因素发生的比值比(OR);Noisy-or 模型假设各风险因素之间相互独立,可以有效地避免不同危险因素之间的相互作用,从而使抑郁的风险预测更加准确。为了克服文献[9—11]中模型的局限性,本文通过对 TILDA 中的数据进行分析,将 Logistic 回归模型与 Noisy-or 模型相结合,建立抑郁症风险预测模型 LRANO,以获得更高的抑郁症早期识别率,从而减少抑郁症的危害。

收稿日期: 2020-10-09; **修回日期:** 2020-11-08

基金项目: 国家自然科学基金(61973149)

第一作者简介: 杨斐(1997—),女,山东烟台人,硕士研究生,研究方向为经济与社会统计。E-mail: yangfei_0422@126.com

通信作者简介: 魏新江(1977—),男,山东东营人,教授,硕士研究生导师,博士,研究方向为非线性系统控制、鲁棒控制等。E-mail: weixinjiang@163.com

1 抑郁症风险预测模型的建立

1.1 Logistic 回归

Logistic 回归分析是一种广义线性回归分析模型[12]。一般情况下,Logistic 回归是用来解决因变量为二元的问题。它不仅可以使用一个或多个预测变量来实现因变量的预测分类,还可以选择对因变量有显著影响的变量[13]。本节采用 Logistic 回归模型确定风险因素的统计显著性,并获得其比值比(OR)。

根据文献[14],导致抑郁症的风险因素主要有五种:性别、残疾、睡眠障碍、丧亲之痛和当前抑郁。本文用 $\delta=(\delta_1,\delta_2,\delta_3,\delta_4,\delta_5)$ 表示这五种风险因素,$\delta_i\in\{0,1,NA\}$,其中 0、1、NA 分别表示第 i 个风险因素不存在、存在和其信息不可得;二元变量 Y 表示未来抑郁状况,$Y\in\{0,1\}$,0 和 1 分别表示无抑郁和抑郁;P 为在风险因素下患有抑郁症的概率,即 $P=P\{Y=1|\delta\}$。建立 Logistic 回归模型如下:

$$\log it(P)=\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1\delta_1+\cdots+\beta_k\delta_k, \quad (1)$$

其中,β_0 为当所有风险因素都不存在时患有抑郁症的风险;$\beta_i(i=1,2,\cdots,k)$ 为第 i 个风险因素的系数,用来解释风险因素的比值比(OR),其计算公式如下:

$$OR_i=e^{\beta_i}。 \quad (2)$$

1.2 Noisy-or 模型

Noisy-or 模型是一种概率图模型,常被用作因果概率模型的一部分[15],通常描述不同风险因素及其共同影响的结果间的相互作用[16]。该模型假设每个原因都会导致结果的发生,且不同的原因对结果的影响是相互独立的。目前,Noisy-or 模型已被广泛用于预测各种疾病的发生概率,如肝病[17]、哮喘[18]和肺恶性肿瘤[19]等。本节将 Noisy-or 模型与 Logistic 回归模型相结合来实现抑郁症的风险预测。

运用 Noisy-or 模型计算未来抑郁的概率取决于风险因素信息的暴露情况。为了更好地实现对未来抑郁状况的预测,基于 Noisy-or 模型建立如下的抑郁症风险评分模型[11]:

$$f(\delta)=1-(1-C_0)\prod_{i\in K}(1-C_i)^{\delta_i}\prod_{j\in U}(1-p_jC_j), \quad (3)$$

式中:p_j 为未知风险因子 j 的患病率,由缺失值预处理后得到的结果代替;K,U 分别为风险因素信息的已知暴露和未知暴露;C_0 是无风险因素存在时未来抑郁的概率,C_i 为在风险因子 i 下未来出现抑郁的条件概率,其表达式为:

$$P(y_i=1|\delta_i)=\begin{cases}0, & \delta_i=0,\\ C_i, & \delta_i=1。\end{cases} \quad (4)$$

模型(3)的基本思想是任何风险因素都可能在未来导致抑郁。换句话说,风险因素对最终抑郁结果的影响是独立的,当且仅当所有风险因素都不存在时,未来患抑郁症的概率为零。此外,除了以上五种风险因素外,本文还考虑其他因素对抑郁症的影响,其概率贡献为式(3)中的 C_0[20]。

根据文献[21]中采用的方法,利用 Logistic 回归中风险因素系数所得到的 OR_i 值来求出条件概率,其计算公式如下:

$$P(y_i=1|\delta_i)=C_i=C_0\frac{OR_i-1}{1-C_0+C_0OR_i}, \quad (5)$$

其中,$C_0=0.061$ 来自文献[22],风险因素丧亲的患病率来自文献[23]。

将计算得到的 C_i 值和其他参数值代入式(3),得到受试者未来抑郁的风险评分。

2 数据描述

2.1 数据集

为了检验由 Logistic 回归模型与 Noisy-or 模型结合构造的抑郁症风险预测模型——LRANO 对抑郁症的预测性能,本文利用爱尔兰老龄化纵向研究数据库(TILDA)对其进行验证。TILDA 数据库是在对爱尔兰老年人进行全面、准确地了解后收集的数据。自 2009 年以来,该数据库每两年收集一次数据,是地方政府制定有关老年人健康、医疗、社会和经济政策的重要依据。在 TILDA 数据库中,共调查了8504名 50 岁以上的受试者,共有1585个变量,其中包括 ID、年龄、认知状况、经济状况、健康状况、收入、生活方式、心理状况、社会参与状况等。

TILDA 数据库中的数据是通过计算机辅助个人访谈(CAPI)技术和自我概念问卷(SCQ)收集

到的,所以原始数据中的变量有很多形式,且存在少量缺失值。为了避免变量的不同形式和缺失值的存在给分析带来困难,需要对原始数据进行预处理。

2.2　数据集的预处理

通过对抑郁症风险因素的相关文献研究发现,性别、残疾、睡眠状况、丧亲和目前的抑郁状态是影响未来抑郁的最重要因素。下面对 TILDA 数据库中的变量进行重编码,得到风险因素的分布描述,如表 1 所示。

由表 1 可以看出,TILDA 数据库中睡眠状况、丧亲、目前抑郁状况和预测变量抑郁中均存在缺失值,其中丧亲变量完全缺失。缺失值的存在会影响最终的预测精度,所以缺失值处理是风险预测模型中的关键问题。

表 1　TILDA 数据库中风险因素及结果的患病率,括号内为 95%置信区间

Tab.1 The prevalence of risk factors and outcomes in the TILDA database with a 95% confidence interval in brackets

风险因素	True	False	Unknown
女性	0.532 2(0.514 1,0.550 3)	0.467 8(0.449 7,0.485 9)	0.000 0
残疾	0.082 8(0.072 8,0.092 8)	0.917 2(0.907 2,0.927 2)	0.000 0
睡眠障碍	0.393 2(0.375 5,0.410 9)	0.606 4(0.588 7,0.624 1)	0.000 4(-0.003 3,0.001 0)
丧亲	0.000 0	0.000 0	1.000 0
当前抑郁	0.257 0(0.241 2,0.272 9)	0.729 0(0.712 8,0.745 1)	0.014 0(0.009 8,0.018 3)
抑郁	0.239 6(0.224 1,0.255 0)	0.743 7(0.727 8,0.759 5)	0.016 7(0.012 1,0.021 4)

目前,缺失值的问题考虑甚少,大多数研究对于缺失值的处理方式往往是直接删除。因此,为了提高预测的准确性,有必要对缺失值进行相应的处理。在本文中,LRANO 模型使用 K 最近邻算法(KNN) 来填充 TILDA 数据库中缺失的值: 对于睡眠障碍和当前抑郁这两个风险因素,由于其缺失的比例非常小,本文在特征空间中寻找 10 个最相似的样本(即最近的邻居) ,通过计算它们的加权平均值来进行填充; 至于风险因素丧亲,TILDA 数据库中没有调查这一变量,本文利用文献 [23] 中报道的患病率 0.041 来代替; 对于预测变量“抑郁”,本文选择直接删除包含该变量缺失值的受试者信息。

3　实验验证及结果

基于风险评估工具 DRAT-up 的风险预测模型(LRANO) ,已在爱尔兰老龄化纵向研究(TILDA) 数据库中得到验证。本节主要介绍如何利用 TILDA 数据库对 LRANO 模型进行验证,以及使用哪些评价指标对模型的性能进行综合评价。

经过上述数据预处理后,仅剩下2922份有效样本。在本文的研究中,第一次数据的调查时间为 2009—2011 年,最终输出结果的调查时间为 2012—2013 年。

接下来,为了检验 TILDA 数据库中风险因素的统计显著性,本文利用 TILDA 数据库中 2009—2011 年的相关数据建立 Logistic 回归模型:

$$\log it(P) = -1.996\,7 + 0.327\,7\delta_1 + 0.754\,2\delta_2 + 0.357\,1\delta_3 + 1.328\,7\delta_5。 \quad (6)$$

通过对各风险因素系数的显著性检验,发现各系数在显著性水平为 0.05 时效果显著。根据式(2) 和式(5) 得各危险因素的 OR 值和 C_i 值,如表 2 所示。

表 2　风险因素的比值比(OR) 、95%的置信区间以及概率贡献值

Tab.2 Odds ratio (OR) of risk factors,95% confidence interval and probability contribution values

风险因素	OR(置信区间)	C_i
性别	1.39(1.1,1.7)	0.023 2
残疾	2.13(1.5,2.9)	0.064 5
睡眠状况	1.43(1.1,1.8)	0.025 6
丧亲	3.3(1.7,4.9) [14]	0.124 0
目前抑郁状态	3.78(3.0,4.8)	0.145 0

基于表 2 中获得的 C_i 值和数据中每个受试者不同风险因素的存在情况,使用风险评分模型(3) 计算每个受试者未来的抑郁状态。为了评估本文风险预测模型的预测性能,本文选择受试者工作特征曲线(ROC) 、AUC 值和平均绝对误差(MAE) 作为性能评价指标。这里,ROC 曲线和 AUC 值用来比较不同模型的预测效果,得出抑郁症的最优分类阈值; 平均绝对误差用来反映预测

值与真实值之间的偏差。采用风险预测模型(3)计算抑郁预测值后,将各受试者抑郁状态的预测结果与实际情况进行比较,计算预测结果的平均绝对误差(MAE),即

$$\varepsilon_{\text{mae}} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|, \quad (7)$$

式中:f_i 为根据风险预测模型(3)计算出的抑郁症预测值;y_i 是受试者抑郁的实际情况;n 是样本总数。

最后,在TILDA数据集中,将LRANO模型与Cattelani等[11]开发的DRAT-up工具进行比较,得到表3;同时,将各受试者的预测值与实际值进行比较,得到ROC曲线图(图1)。

表3 LRANO和DRAT-up在TILDA数据库中的性能比较

Tab.3 Performance comparison between LRANO and DRAT-up in TILDA database

评价指标	LRANO	DRAT-up
AUC	0.731 3	0.712 0
MAE	0.288 7	0.292 9

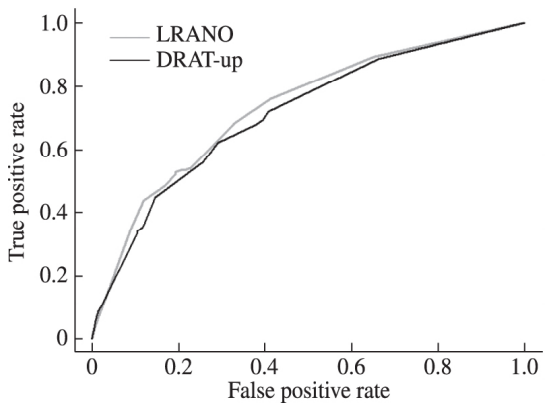

图1 LRANO和DRAT-up工具在TILDA数据库中的ROC曲线

Fig.1 ROC curves of LRANO and DRAT-up in TILDA database

由表3可以看出:在TILDA数据库中DRAT-up的AUC值为0.712,平均绝对误差为0.292 9;LRANO的AUC值为0.731 3,平均绝对误差为0.288 7。显然,LRANO模型的AUC值比DRAT-up高0.019 3,平均绝对误差比DRAT-up低0.004 2,表明LRANO的预测性能优于DRAT-up。结合图1,LRANO模型的ROC曲线位于DRAT-up模型ROC曲线的右上方,进一步说明LRANO的预测性能优于DRAT-up。

4 结语

本文主要讨论抑郁症的早期识别和预测问题,将Logistic回归模型与Noisy-or模型相结合,得到一个新的LRANO模型。该模型克服现有抑郁症风险预测模型预测精度低、方法复杂等缺陷,在TILDA数据库中实现LRANO模型的可行性。由于获取数据困难,对于LRANO模型的进一步研究应考虑在不同的数据集中进行验证。

参考文献:

[1] 舒敏.南昌市高校学生抑郁症状况及影响因素调查[D].南昌:江西财经大学,2017.

[2] OOI K E B,LECH M,ALLEN N B. Multichannel weighted speech classification system for prediction of major depression in adolescents[J].IEEE Transactions on Biomedical Engineering,2013,60(2):497-506.

[3] KINTZIGER M C W.Late life depression:a global problem with few resources[J].Psychiatric Clinics of North America,2013,36(4):475-481.

[4] LUPPA M,SIKORSKI C,MOTZEK T,et al. Health service utilization and costs of depressive symptoms in late life—a systematic review[J].Current Pharmaceutical Design,2012,18(36):5936-5957.

[5] BELVEDERI M M,AMORE M,RESPINO M,et al.The symptom network structure of depressive symptoms in late life:results from a European population study[J]. Molecular Psychiatry,2020,25(7):1447-1456.

[6] BEKERIS J,WILSON L A,FIASCONARO M,et al. New onset depression and anxiety after spinal fusion surgery:incidence and risk factors[J].Spine,2020,45(16):1161-1169.

[7] TANG H Y,LI N,MAO J W,et al.Depression and its risk factors in inpatients with gastrointestinal diseases in department of gastroenterology of general hospital[J].Cancer Gene Therapy,2007,4(4):364-371.

[8] HUANG Y,SU Y H,JIANG Y,et al.Sex differences in the associations between blood pressure and anxiety and depression scores in a middle-aged and elderly population:The Irish Longitudinal Study on Ageing(TILDA)[J].Journal of Affective Disorders,2020,274:118-125.

[9] SHIN K R,JUNG D,JO I,et al. Depression among community dwelling older 10 adults in Korea:a prediction model of depression[J].Archives of Psychiatric Nursing,2009,23(1):50-57.

[10] OKAMOTO K,HARASAWA Y.Prediction of symptomatic depression by discriminant analysis in Japanese community-dwelling elderly [J].Archives of Gerontology and Geriatrics,2011,52(2):177-180.

[11] CATTELANI L,MURRI M B,CHESANI F,et al.Risk prediction model for late life depression: development and validation on tree large European datasets [J].IEEE Journal of Biomedical and Health Informatics,2019,23(5):2196-2204.

[12] 杨英茹,黄媛,高欣娜,等.基于 Logistic 回归模型的设施番茄病毒病预警模型构建 [J].河北农业科学,2019,23(5):91-94.

[13] CHEN B J,CHEN X F,LI B,et al.Reliability estimation for cutting tools based on logistic regression model using vibration signals [J].Mechanical Systems and Signal Processing,2011,25(7):2526-2537.

[14] COLE M G,DENDUKURI N.Risk factors for depression among elderly community subjects: a systematic review and meta analysis [J].American Journal of Psychiatry,2003,160(6):1147-1156.

[15] 张俊光,徐振超,贾赛可.基于 Noisy-or Gate 和贝叶斯网络的研发项目风险评估方法 [J].科技管理研究,2015,35(1):193-196.

[16] CHENG J Z,ZHU C,FU W L,et al.An limitation medical diagnosis method of hydro-turbine generating unit based on bayesian network [J].Transactions of the Institute of Measurement and Control,2019,41(12):3406-3420.

[17] ONISKO A,DRUZDZEL M,WASYLUK H.Learning bayesian network parameters from small data sets: application of Noisy-OR gates [J].International Journal of Approximate Reasoning,2001,27(2):165-182.

[18] ANAND V,DOWNS S M.Probabilistic asthma case finding: a noisy or reformulation [J].American Medical Informatics Annual Symposium,2008:6-10.

[19] LIAO F Z,LIANG M,LI Z,et al.Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network [J].IEEE Transactions on Neural Networks and Learning Systems,2019,30(11):3484-3495.

[20] 胡勇健,肖志怀,周云飞,等.基于贝叶斯网络 Noisy Or 模型的水电机组故障诊断研究 [J].水力发电学报,2015,34(6):197-203.

[21] GUNTHER E.FRAT-up,a web-based fall-risk assessment tool for elderly people living in the community [J].Journal of Medical Internet Research,2015,17(2):e41.

[22] JOHN R M,BEEKMAN A T F,BRAAM A W,et al.Depression among older people in Europe: the Europe studies [J].World Psychiatry: Official Journal of the World Psychiatric Association,2004,3(1):45-49.

[23] TURVEY C L,CARNEY C,ARNDT S,et al.Conjugal loss and syndromal depression in a sample of elders aged 70 years or older [J].American Journal of Psychiatry,1999,156(10):1596-1601.

A Risk Prediction Model for Depression Based on Logistic Regression and Noisy-or

YANG Fei, WEI Xinjiang

(School of Mathematics and Statistics Science,Ludong University,Yantai 264039,China)

Abstract: By combining the Logistic regression model with the Noisy-or model,a risk prediction model LRANO was proposed for the early identification and detection of depression.In this model,the Logistic regression model was taken to calculate the probabilistic contribution of different risk factors to depression,and the Noisy-or model was combined to integrate the various model parameters to form the final risk prediction model for depression.In addition,the AUC value of the model is 0.731 3 and the average absolute error is 0.288 7 through validation on The Irish Longitudinal Study on Aging(TILDA),which indicates the validity of the model.

Keywords: depression; Logistic regression; Noisy-or; risk prediction

(责任编辑　顾建忠)