



新时代 积极应对人口老龄化 研究文集·2019

中国老年学和老年医学学会 主编

Collected Papers of "Actively
Responding to Population Ageing
in the New Era"



华龄出版社
HUALING PRESS

新时代
积极应对人口老龄化
研究文集·2019

Collected Papers of "Actively
Responding to Population Ageing
in the New Era"

ISBN 978-7-5169-1474-8



9 787516 914748 >

定价：88.00元

责任编辑：徐 芳

责任印制：李未圻

图书在版编目 (CIP) 数据

新时代积极应对人口老龄化研究文集 . 2019 / 中国老年学和老年医学学会主编 . —北京：华龄出版社，2019.10
ISBN 978-7-5169-1474-8

I . ①新… II . ①中… III . ①人口老龄化—中国—文集 IV . ① C924.24-53

中国版本图书馆 CIP 数据核字 (2019) 第 224174 号

书 名：新时代积极应对人口老龄化研究文集 . 2019
作 者：中国老年学和老年医学学会

出 版 人：胡福君

出版发行：华龄出版社

地 址：北京市东城区安定门外大街甲 57 号

邮 编：100011

电 话：010-58122246

传 真：010-84049572

网 址：<http://www.hualingpress.com>

印 刷：天津中印联印务有限公司

版 次：2019 年 10 月第 1 版 2019 年 10 月第 1 次印刷

开 本：787 × 1092mm 1/16

印 张：26.5

字 数：500 千字

定 价：88.00 元

版权所有 翻印必究

本书如有破损、缺页、装订错误，请与本社联系调换

目录

CONTENTS

第一篇 国庆 70 周年特稿

新中国 70 年人口老龄化发展趋势与原因分析	2
------------------------------	---

第二篇 促进老年健康：理论与实践

我国老年人健康状况、问题及对策研究	24
老年疾病的特点与防治	31
老年病防治工作建设	37
老年健康教育刍议	43
以转变生活方式促进健康老龄化	50
政府、社会、家庭在老年健康促进中的作用	57
中老年慢病患者的健康信息素养现状及影响因素	61
基于逻辑回归的中老年人持续使用在线健康社区预测研究	73
我国老年人体育锻炼的行为特征、支持及影响因素研究 ——基于 2016 年 CLASS 调查数据	82
推动老年健康的难点和重点在农村	92
农业转移人口市民化对老年健康的影响 ——基于 CLHLS 数据的实证研究	97
不同养老模式老年人心理健康状况调研	116
养老院老年人锻炼心理需求量表的跨文化调适	125

基于逻辑回归的中老年人持续使用 在线健康社区预测研究^{*}

郭鑫鑫 左美云 许洁萍

(中国人民大学信息学院 智慧养老研究所)

摘要：国家积极倡导“互联网+健康养老”，在政策的鼓励和资本的支持下，在线健康社区出现并被很多人使用，其对患有慢性病的老年人的自我管理和疾病控制具有积极影响。用户的持续使用是在线健康社区存在和发展的基础，因此本文使用逻辑回归算法对甜蜜家园论坛中的中老年用户建立持续使用预测模型，探索影响中老年用户持续使用的重要因素，从而给出运营建议。

关键词：逻辑回归；在线健康社区；持续使用

一、引言

2019年7月15日，《国务院关于实施健康中国行动的意见》(以下简称《意见》)提出一系列提高全民健康水平的专项行动，面向孩子、老人、劳动者等重点人群，覆盖心脑血管疾病、癌症、糖尿病等疾病。国家卫生健康委发布的《健康中国行动(2019—2030年)》中鼓励和支持“互联网+健康养老”，充分利用信息技术做好慢性病健康管理，同时指出患有慢性病的老年人应主动获取自我管理的知识和技能，做好自我管理，延缓病情发展，减少并发症。

在政策的积极引导和资本的鼓励下，互联网健康医疗行业已经形成了包括健康服务、精准医疗和医药研发等在内的健康医疗大数据产业链。其中，在线健康社区(Online Health Community, OHC)逐步出现并被很多人所使用，在线健康社区分别提供了医患之间、患患之间、医医之间的交流平台，其交流互动功能对于慢性病患者的自

^{*} 基金项目：2019年教育部哲学社会科学研究重大课题攻关项目(智慧化养老服务研究)、国家自然科学基金资助项目(71771210)、北京市自然科学基金资助项目(9182008)。

我健康管理、疾病控制有着积极影响 (Martijn et al, 2013)。OHC 用户的持续使用是 OHC 生存和发展的基础,那么,哪些因素会影响用户的持续使用?本文将采用逻辑回归算法对 OHC 的中老年用户的持续使用行为进行预测建模,探索影响中老年用户持续使用的关键因素,并针对如何促进中老年用户活跃且持续使用提出建议,从而帮助中老年人从 OHC 中获益,更好地实现自我管理。

二、相关研究及方法

(一) 文献回顾

在线健康社区是信息系统 (Information System, IS) 的一种应用方式,学者们对信息系统的持续使用做了大量的研究。用户的持续使用行为受到多种因素的影响,已有研究应用不同理论和模型从使用动机和影响因素角度来讨论用户对于不同信息系统的持续使用行为,应用到的理论有期望确认理论、信息系统成功模型、社会支持理论等。对在线健康社区的研究中,在用户使用动机方面,翟羽佳等 (2017) 使用“百度戒烟吧”的用户数据,根据用户的停留时间分成短期用户和长期用户,并分析出短期用户的参与动机主要是寻求社会支持和自我满足,长期用户扮演欢迎者的角色,是利他主义的执行者。在影响因素的研究中,张星等 (2016) 发现在线健康社区系统质量、信息质量显著影响用户满意度,评价支持和情感支持显著影响用户的社区归属感,用户的满意度和社区归属感共同决定用户的忠诚度。

逻辑回归 (Logistic Regression, LR) 是一种对样本分类预测的常用算法,它可以建立因变量与多种自变量关系的模型,拥有良好的分类效果。Qiu et al (2011) 对在线癌症患者社区中的帖子使用逻辑回归等算法进行情绪分析,最终得出情绪模型来预测参与者的情绪及情绪动态变化过程;牟冬梅等 (2016) 使用逻辑回归及其他算法对电子病历系统中的孕妇电子病历数据进行研究,从而得出针对妊高症诊断的预测模型,并筛选出影响妊高症的危险因素有年龄、体重、收缩压等。

对于在线健康社区的持续使用研究,学者们较少使用逻辑回归算法等机器学习方法对客观网络数据进行分析和挖掘,根据逻辑回归算法在信息系统特别是在 OHC 的广泛应用,在本研究中,将采用逻辑回归算法对 OHC 中的用户数据建立持续使用预测模型,并分析影响用户持续使用 OHC 的关键因素。

(二) 逻辑回归基本原理

逻辑回归 (Logistic Regression, LR) 是机器学习中的一种常用算法。数据集包含

自变量 x_i 和因变量 y_i 。该算法假设样本分类变量即因变量 y_i 服从伯努利分布, y_i 的取值只有两个: 0 或 1, 即用 $y_i \in \{0,1\}$ 来分别表示样本的两个分类, 即正类和负类 (唐亘, 2018)。

(三) 评价标准

本研究采用十折交叉验证, 引入正确率、准确率 (Precision, 也称查准率)、召回率 (Recall, 也称查全率)、F1 值等标准来对模型的性能进行估计。

三、持续使用预测模型

本研究中的持续使用是指在信息系统领域, Bhattacharjee (2001) 最初提出 IS 持续使用模型 (Expectation confirmation model of IS continuance, ECM-ISC), 认为持续使用是指用户初始采纳 IS 后并未中断使用的持续意愿和行为, 即个体继续使用信息系统的主观倾向。该模型成为揭示 IS 用户持续使用行为规律的最具代表性的理论。Moez Limayem et al (2007) 认为 IS 持续使用及持续使用习惯描述了一种继续使用某一 IS 的行为模式。

因此我们认为, 在 IS 领域, 持续使用指的是一种采纳某 IS 后未改变使用意愿的行为方式。在本研究中将以未来一个月内用户是否登录在线健康社区为基准来判断用户是否持续使用, 用 y_i 表示, $y_i \in \{0,1\}$, 用 x_i ($i=1 \cdots n$) 表示用户的特征值。

假设函数: $h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

其中 θ_0 表示偏置, 或称截距; $\theta_1 \cdots \theta_n$ 表示对应特征的权值。

逻辑回归函数: $S_{\theta}(x) = 1/(1 + e^{-\theta^T x})$

于是, 预测模型为: $\hat{y}_i = \begin{cases} 1, & S_{\theta}(x) > 0.5 \rightarrow \hat{h}_{\theta}(x) > 0 \\ 0, & \text{else} \end{cases}$

该模型的损失函数为: $L(\theta) = -\frac{1}{n} \sum_i y^{(i)} \log(\hat{h}_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - \hat{h}_{\theta}(x^{(i)}))$

通过求解损失函数的最小值即得到一组特征的权值和截距, 从而得出预测模型。

四、数据来源与处理

“甜蜜家园论坛”搭建于 2005 年 9 月 8 日, 是中国最早的专注于糖尿病的在线健康社区之一, 目前拥有大约 440000 名用户, 具有健康资讯分享和在线交流等功能, 用户量较大且较为活跃, 故本研究以“甜蜜家园论坛”为数据来源。

(一) 数据筛选

使用自主编写的爬虫程序对“甜蜜家园论坛”的用户数据进行采集，于2019年4月25日采集了从2005年9月8日到2019年4月25日的部分用户数据共38990条。

原始数据包括用户基本信息、用户历史操作行为和基本统计信息共计22个字段。用户基本信息包含用户ID、性别、出生日期等，用户历史操作行为包含回帖数、发帖数等字段，基本统计信息包含好友数、空间访问量等字段。

由于数据集稀疏，且在本研究中用户年龄是重要特征，故筛选出“出生日期”字段不为空的用户数据共698条。由于本研究重点研究中老年人的持续使用行为，故根据普遍的人类年龄段划分方法，选取45岁及以上的用户数据共226条作为最终的数据集。部分数据展示如图1所示。其中，用户是否持续使用由2019年5月26日采集的“用户上次活跃时间”字段来判断。

userid	username	zoneid	status	state	friends	postcount	topiccount	tribetype	usersex	userbirthdate	userage	userplace	isigncc	continues	thismonth	lastsigntime	userrank	usertype	allonlinefine
5305	无花果	13506	未验证	20	8633	364	其他	保密	1969年2	50	12	1	1	2019-4-19 16:45	3			2007-3-1 8:55	5963小时
11870	虞美人	8865	未验证	8	3147	339	X型	保密	1969年1	50	20	2	0	2016-4-7 7:01	4			2007-11-16 23:07	1099小时
3305	梦想天使	6444	未验证	24	5888	367	患者	女	1972年4	47								2006-10-14 9:02	6222小时
6833	看看云	4019	未验证	14	1916	99	患者	男	1961年6	58								2007-5-11 14:07	6381小时
12078	2002xldg	3071	未验证	5	1294	146	2型	保密	1957年2	62								2007-11-24 19:36	4794小时
2607	柳絮飞扬	1838	未验证	3	1424	202	2型	男	1970年8	49								2006-8-16 13:47	4846小时
29831	惠州平平	1817	未验证	44	146	245	2型	保密	1956年5	63	9	1	0	2018-6-18 21:20	3			2008-11-13 8:32	1925小时
662	罗氏	1574	未验证	9	1027	7	患者	保密	1968年1	61	1	1	0	2017-1-16 13:13	1			2006-1-9 21:03	4943小时
572	yuxiguī	1113	未验证	0	323	16	患者	男	1960年4	59								2005-12-23 14:48	927小时
1125	飞鸟未尽	849	未验证	4	721	49	2型	男	1965年8	54								2006-3-10 10:36	194小时

图1 部分数据展示

(二) 数据预处理

首先我们去掉了数据集里缺失值超过50%的字段如“居住地”“累计签到总天数”等，然后对其余字段中的少量缺失值进行补全，如将“性别”中的缺失值填充为“保密”等。

在机器学习中算法输入的变量也称特征，数据集中包含的226名用户中有142名男性、21名女性及63名性别未知，用保密替代。由于“用户类型”和“性别”两个特征的特征值是无序的，所以我们将这两个特征进行one-hot编码(Audreas et al, 2018)处理，one-hot编码将特征的n个可能的取值变换为长度为n的二进制特征向量，其中只有一个位置是1，其余位置是0。one-hot编码会增加数据维度，会使特征矩阵更加

稀疏，所以为减少数据稀疏性，“邮箱验证状态”“上次活跃时间点”和“注册时间点”特征仅作了整数编码（integer encoding）处理。我们构造了用户间隔天数、日均活跃度、相对活跃度等特征，最终形成了 15 个特征，具体解释如表 1 所示：

表 1 用户特征表

特征名称	备注
空间访问量	指用户的虚拟信息展示网页被其他用户访问的次数
邮箱验证状态	指用户的邮箱是否已验证有效
好友数	指用户在论坛中的好友数量
回帖数	指用户回复他人帖子的数量
发帖数	指用户发布帖子的数量
年龄	用户年龄
在线时间	用户从注册到数据采集时的在线时长，单位为小时
积分	指用户在论坛规则下获取的积分
上次活跃时间点	类别型变量，包括上午（6—12）、下午（12—18）、晚上（19—24）、凌晨（0—6）
注册时间点	同上
间隔天数	指用户注册时间和上次活跃时间之间的间隔天数
日均活跃度	等于（发帖数 + 回帖数）/ 间隔天数
相对活跃度	等于积分 / 间隔天数
类型	包括患者、家属、客服、商家、医生五种类型，其中客服指论坛运营人员，商家指在论坛投放广告的广告主
性别	包括男、女、保密三种类型

五、实验结果与分析

（一）关键特征和运营建议分析

使用 Python 和 scikit-learn 库中的逻辑回归算法处理后，并分析特征系数的显著度，最后得出显著特征的系数排序结果和截距值如表 2 所示。我们可以看到，距离用户注册论坛的时间越久，用户越会持续使用论坛；此外，用户年龄、男性、类型为患者、

用户积分等是非常重要的影响因素，相比较而言，用户注册时间点和上次活跃时间点等特征的影响程度不高。用户的日均活跃度系数为负值，结合数据集得知，有一些用户从注册到离开论坛的间隔天数短但是回帖较多，导致该类用户日均活跃度变大，这些离散点导致了这一特征的系数为负值。

表2 特征值系数排序和截距值

特 征	系 数	特 征	系 数
间隔天数	2.809	相对活跃度	-0.231
年龄	0.785	性别—保密	0.226
性别—男	-0.614	主题数	0.164
用户积分	-0.461	空间访问量	0.128
类型—患者	0.458	性别—女	0.124
日均活跃度	-0.448	类型—医生	-0.106
邮箱验证状态	-0.446	类型—家属	0.026
在线时间	0.348	上次活跃时间	0.022
好友数	0.336	注册时间	0.009
回帖数	0.308	类型—客服	-0.001
类型—商家	-0.276	截 距	-3.710

基于以上分析，我们可以得出以下结论：

(1) 注册时间久的用户，有较大的概率发生持续使用行为，所以对于论坛的运营来说，不仅需要关注发展新用户，更应该多多关注老用户。

(2) 在当前的论坛积分奖励规则下，即总积分 = 精华帖数 × 10 + 威望 × 50 + 金钱，其中威望包含用户的精华帖数和排行榜排名，金钱可通过发帖、回帖和签到等方式获得。用户的积分不会促进用户持续使用，运营者可以适当调整积分奖励规则，如减少积分发放个数或增加积分获取难度，然后对积分排名前一百的用户进行奖励，使用户之间形成一定的竞争关系，以此增加用户的论坛持续使用率。

(3) 用户的好友数和回帖数对用户持续使用行为有积极的促进作用，故论坛运营人员可以采取一定措施鼓励用户互相加好友和回帖，如可以发布精华帖或发布活动话题，吸引用户参与回帖和发帖等。

(二) 算法性能

我们比较了原始的特征体系和经过 one-hot 编码处理后的特征体系下的逻辑回归算法的性能,分别采用十折交叉验证,结果如表 3 所示,可以看到,经过 one-hot 编码处理后的算法性能大大提高,在正确率、准确率、召回率和 F1 值等标准上都有大幅的改进。

表 3 不同特征体系下的逻辑回归算法性能

特征体系 + 算法	正确率 /%	准确率 /%	召回率 /%	F1-Score/%
原始特征体系 +LR	91.2	75	60	66.7
one-hot 编码后的特征体系 +LR	97.1	87.5	87.5	87.5

六、结束语

本文利用爬虫程序获取了在线健康社区中的客观网络数据,并通过逻辑回归算法根据用户的基本信息和历史行为数据构建出用户的持续使用预测模型,其中对分类特征进行了独热编码处理,大幅度提高了逻辑回归算法的性能。根据最佳的特征系数分析了影响用户持续使用在线健康社区的关键因素,为在线健康社区的平台运营等提供了参考建议,使用户保持更高的活跃度,使平台持续健康发展。本文的不足之处在于仅使用了逻辑回归算法,在特征选择上构建的特征不够多,而且没有考虑好的办法来解决离散点较多的问题,今后可考虑多种机器学习及融合算法,并使用多种数学关系来构建和选择特征。

参考文献

- [1] 《国务院关于实施健康中国行动的意见》,新华网,2019年7月15日。
- [2] 《健康中国行动(2019—2030年)》,卫生健康网站,2019年7月15日。
- [3] 《2018年大数据时代下的健康医疗行业研究报告》,艾瑞咨询,2018年5月。

[4] van der Eijk Martijn, Faber Marjan J, Aarts Johanna W M, Kremer Jan A M, Munneke Marten, Bloem Bastiaan R. Using online health communities to deliver patient-centered care to people with chronic conditions. [J]. Journal of medical Internet research, 2013, 15(6).

[5] 刘萌萌、邓朝华:《在线健康社区用户参与行为综述》,《医学信息学杂志》2018年第11期。

[6] 翟羽佳、张鑫、王芳:《在线健康社区中的用户参与行为——以“百度戒烟吧”为例》,《图书情报工作》2017年第7期。

[7] 张星、陈星、夏火松等:《在线健康社区中用户忠诚度的影响因素研究:从信息系统成功与社会支持的角度》,《情报科学》2016年第3期。

[8] Qiu B, Zhao K, Mitra P, et al. . Get online support, feel better—sentiment analysis and dynamics in an online cancer survivor community [M]. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and Third International Conference on Social Computing, 2011: 274 - 81.

[9] 牟冬梅、任珂:《三种数据挖掘算法在电子病历知识发现中的比较》,《现代图书情报技术》2016年第6期。

[10] Ba S, Wang L. Digital health communities: The effect of their motivation mechanisms[J]. Decision Support Systems,2013, 55(4):941 - 947.

[11] 唐焜:《精通数据科学:从线性回归到深度学习》,人民邮电出版社,2018年版。

[12] Bhattacharjee A.Understanding Information:An Expectation-Confirmation Model[J]. Mis Quarterly, 2001, 25 (3) :351-370.

[13]Limayem M, Hirt S G, Cheung C M K.How Habit Limits the Predictive Power of Intention:The Case of Information Systems Continuance[J].MIS Quarterly, 2007, 31 (4) :705-737.

[14]Gooden R J, Winefield H R. Breast and Prostate Cancer Online Discussion Boards: A Thematic Analysis of Gender Differences and Similarities[J]. Journal of Health Psychology, 2007, 12(1): 103-114.

[15]Audreas C. Muller、Sarah Guido[M]. 张亮译:《Python 机器学习基础教程》,人民邮电出版社,2018年版。

[16]Ian Goodfellow, Yoshua Bengio,Aaron Courville.Deep Learning[M]MITP Verlags GmbH, 2018.

[17] 张星、夏火松、陈星、侯德林:《在线健康社区中信息可信性的影响因素研究》,《图书情报工作》2015年第22期。

[18] 吴江、李姗姗:《在线健康社区用户信息服务使用意愿研究》,《情报科学》2017年第4期。

[19] 邓朝华、莫秀婷:《移动健康服务用户使用意愿的实证分析》,《中国卫生统计》

2015年第5期。

[20] 赵栋祥：《国内在线健康社区研究现状综述》，《图书情报工作》2018年第9期。

[21] 吴菊华、王煜、黎明、蔡少云：《基于加权知识网络的在线健康社区用户知识发现》，《数据分析与知识发现》2019年第2期。

作者简介

左美云，中国人民大学信息学院，教授。