Interdisciplinary Science Letters 交叉科学快报, 2023, 7(1), 17-23
Published Online March 2023 in Hans. https://www.hanspub.org/journal/isl
https://doi.org/10.12677/isl.2023.71004

Hans汉斯

基于机器学习的国家认同的影响因素研究

王宇拓，侯牧天，唐　燊

西南交通大学应用心理学研究院，四川 成都

收稿日期：2023年2月11日；录用日期：2023年3月6日；发布日期：2023年3月15日

摘　要

本文基于2019年中国社会状况综合调查(CSS2019)，探索国家认同的影响因素。使用数据清洗后的266个数据变量建立MLP模型，来预测被试的国家认同，使用DALEX包来探索最具价值的预测因子。总体来说，机器学习模型确定的预测因子与现有的研究结论是一致的，例如：经济因素、宗教、社会保障等。同时，本文补充了前人对于国家认同的影响因素的研究，为“对法治的感受”可影响国家认同提供了数据支撑。我国应全面推进依法治国，加快建设社会主义法治国家。同时，本文也是将机器学习技术应用于心理学领域的一次积极尝试。

关键词

国家认同，开放数据，机器学习

Research on the Influencing Factors of National Identity Based on Machine Learning

Yutuo Wang, Mutian Hou, Shen Tang

Institute of Applied Psychology, Southwest Jiaotong University, Chengdu Sichuan

Received: Feb. 11th, 2023; accepted: Mar. 6th, 2023; published: Mar. 15th, 2023

Abstract

Based on the 2019 Chinese Social Survey (CSS 2019), this research explores the influencing factors of national identity. After data preprocessing, we use the 266 variables to build an MLP model to predict the national identity of the subjects and use the DALEX package to select the most valuable predictor. In general, the predictive factors determined by the machine learning model are consistent with results of prior studies, such as economic factors, religion, social security, etc. This research supplements previous studies on the factors affecting national identity, and provides data support

文章引用: 王宇拓, 侯牧天, 唐燊. 基于机器学习的国家认同的影响因素研究[J]. 交叉科学快报, 2023, 7(1): 17-23.
DOI: 10.12677/isl.2023.71004

for "feelings of the rule of law" that can affect national identity. China should comprehensively promote the rule of law and accelerate the construction of a socialist country ruled by law. At the same time, this research is an active attempt to apply machine learning technology to the field of psychology.

Keywords

National Identity, Open Data, Machine Learning

1. 引言

国家认同对于国家的经济发展、社会稳定、民族团结、国家统一都具有重要意义。国际上，全球化、多元化、多样化是当今世界的主要趋势。而中国是一个幅员辽阔的多民族、多语言、多文化的国家，人们存在不同程度的国家认同层次以下的民族认同、地域认同和文化认同[1] [2]。目前，我国面临着复杂的国际国内环境，复杂激荡的环境会影响人们的认同观念，在此背景下，探索影响我国国家认同的因素势在必行。

本研究基于 2019 年中国社会状况综合调查(CSS2019)数据来探索国家认同的影响因素，旨在探索更全面的影响因素，为相关单位和政府决策提供依据，另一方面以全体中国人为研究对象的相关研究较少，本研究的研究对象为全国 18~69 周岁的住户。过去研究大都基于 CSS2013，而本研究基于 CSS2019，数据较新，可探索近年来中国社会中关于国家认同的影响因素，并且过去研究者大都采用建立多元线性回归分析模型的方法来进行数据分析，该方法得到的影响因素均与国家认同有着直接的线性关系，本研究通过建立机器学习模型得到的影响因素与国家认同之间可能具有非线性关系和复杂交互作用。

基于 DALEX 包识别 MLP 中的预测因子发现，经济因素会影响国家认同感，具体而言，收入越高国家认同感越弱；“社会保障感知”有利于增加国家认同；“对法治的感受”、“宗教信仰”和国家认同之间存在着复杂的交互作用。

2. 机器学习模型构建

2.1. 数据集

本研究数据集来自 CSS2019，CSS2019 为中国社会质量基础数据库公布的最新数据集。CSS2019 共包含 10,283 行数据，每行数据代表一个被试，问卷有 1016 个数据变量。得益于中国社会科学院社会学研究所对于问卷编制、抽样、执行管理、质量监控等环节的严格把控，由 CSS 数据资料得出的研究结果可推论至全国 18~69 周岁的住户。2019 年的 CSS 调查既包括测量国家认同的数据变量，也包括潜在可预测国家认同的数据变量，因此我们选择该数据集。

2.2. 目标变量

过去研究中，有研究者将 CSS2013 有关国家认同的 5 道题目作为国家认同问卷，根据被试在该问卷上得分高低判断其国家认同的强弱[3]，在 CSS2019 中对应部分仅保留 CSS2013 中的 2 道题目，3 道题目做出

了修改。CSS2019 中题目包括“现在大多数人都没有什么信仰”、“我经常为国家取得的成就而感到自豪”、“如果有下辈子，我还是愿意做中国人”、“每个中国人都有同样的机会获取财富与幸福”和“没有共产党，中国就会陷入混乱”，问卷采用李克特 5 点计分。对这 5 道题目进行一致性检验，其克隆巴赫 α 系数为 0.595。对这 5 道题目进行单因子 CFA 来检验试题质量，CFA 结果显示，第 1 道题目的标准化因子载荷值为 0.184，小于 0.5，说明该项目无法较好地反映所测量的构念[4]，且第 1 道题目并未言明“信仰”是政治信仰或宗教信仰，容易使被试产生误解，从而导致该项目质量不佳，可删去该题项。删去该题项后，各道题目标准化因子载荷均大于 0.5，并且 CFA 拟合指标为 $\chi^2/df = 2.80$，RMSEA = 0.013，CFI = 1.00，TLI = 1.00，SRMR = 0.004，克隆巴赫 α 系数上升至 0.633，各项指标符合心理测量学要求。因此，保留后 4 道题目共同作为国家认同的判断指标，其中被试选择“很同意”的比例分别为 59.20%、76.68%、48.80%、68.01%。本研究将被试的反应由连续变量转换为二分变量，在 10,283 个被试中，有 3288 名(32.0%)被试 4 个题项均选择“很同意”，我们将其编码为“高国家认同”，其余 6995 名(68.0%)被试被编码为“低国家认同”。

2.3. 数据预处理

通过数据集清洗、填补缺失值、使用哑特征、特征缩放后，共保留了 266 个数据变量，最终保留数据 10280 行。

2.3.1. 数据集清洗

我们删除了部分由研究者创建的数据变量，例如：uid、ID、weight 等；删除了一些缺失值过多不能帮助产生预测国家认同假设的题项，例如：问卷 A 部分的家庭成员情况、问卷 B 部分关于家庭耕地等具体情况、问卷 C 部分关于家庭房产的具体情况等(Abhishek *et al.*, 2020)；删除了一些开放性回答的题项，例如：“其他方面的生活压力和困难，请说明___”、“其他群，请说明”、“其他团体(其他，请注明)”等；删除了构成目标变量的 4 道题项。数据清洗后，共保留了 266 个数据变量。

2.3.2. 填补缺失值

CSS2019 存在部分缺失数据，数据清洗后数据集上仍有 32.1%的缺失值。首先通过分析调查问卷逻辑填补部分数据，例如：当被试在题项“您家目前有没有自有住房？”选择为 2 (没有自有住房)时，则将该被试在题项“您家目前有几套自有住房？”的值由缺失填充为 0；对于 g7_1~g7_20 (您认为一个好的社会应包括下列哪些特征？)，分析数据集发现，被试选择了某一项则对应值为 1，否则为缺失，我们可以将缺失值填充为 0，代表被试未勾选此项。然后对于数值型变量例如：“总收入：金额”、“生活消费总支出”以及“家庭总收入”等使用序列平均值进行插补。对于分类变量例如：f4b3_1~f4b3_3 (富人能取得财富最主要的原因是什么？最多选三项，并排序)，部分被试仅选择一个原因作为最主要原因，部分被试选择三个原因并排序，可以将该情况下的缺失作为分类变量的另一个值，可充分反映被试作答情况。此时，数据集的完整度为 99.9%，最后删去仍包含缺失值的个案，最终保留数据 10,280 行。

2.3.3. 哑特征的使用

在数据集中存在部分类别特征，连续输入使得估计器认为类别之间是有序的，但实际上分类值与大小无关，因此把这种类别特征拆分为多个哑特征，例如题项“您的政治面貌是”：选项有 1 (中共党员)、2 (共青团员)、3 (民主党派)、4 (群众)这四类，将其拆分为四个二元分类的特征，答案均为是或否，计算机读入数值 1 或者 0。

2.4. 模型构建

为了训练我们的模型，我们使用 80%的数据作为训练集来训练模型，使用余下 20%的数据作为测试

集来测试模型。本研究旨在构建一个 MLP 神经网络，将样本的特征映射到对应标签上。MLP 神经网络是一种前向结构的人工神经网络，第一层为输入层，中间为隐藏层，最后一层为输出层。隐藏层和输出层每个节点的关系公式如下，式中 I_i 为第 l 层第 i 个节点的输入，O_i 为第 l 层第 i 个节点的输出，w_{ji} 是连接点 j，i 的权重，x_j 是与节点 i 相连接的上一层节点 j 对 i 的输入值，θ_i 是节点 i 的阈值。

$$I_i = \sum_{j=1}^{n} w_{ji} x_j - \theta_i \tag{1}$$

$$O_i = f\left(I_i\right) \tag{2}$$

首先将模型中的隐藏层数量由 0 到 5 逐步增加进行实验(Abhishek *et al.*, 2020)，我们发现，在 MLP 由单隐藏层增加至 2 个隐藏层时，准确度有一定提高，但是增加至 3 个隐藏层时，准确度增加不明显，因此我们选择有 2 个隐藏层的模型。为了确定更好的模型，我们调试了一些超参数，包括隐藏层神经元的数量、批处理数量、学习率，也使用了不同的激活函数如：identity、logistic、tanh、和 relu，使用了不同的优化权重算法如：lbfgs、sgd、adma。在精确率和召回率均大于 50%的模型中选择准确率最高的模型，表 1 给出了我们最终模型中使用的参数。

Table 1. MLP model parameters
表 1. MLP 模型参数

Parameter	Value
Hidden_layer_sizes	(26,61)
Activation	relu
Solver	adam
Batch size	200
Learning rate	0.001
Max_iter	200
Random_state	50

注：Hidden_layer_sizes 表示隐藏层的大小；Activation 表示激活函数；Solver 表示权重优化算法；Batch size 表示批尺寸；Learning rate 表示学习率策略；Max_iter 表示算法收敛的最大迭代次数；Random_state 表示随机种子的数量。

3. 结果

3.1. 国家认同的可预测性

在测试集(未参与模型构建的数据集)中，MLP 模型准确地将 69.8% (95%CI = [67.8%, 71.8%])的被试识别为高国家认同或低国家认同。表 2 给出了 MLP 模型使用测试集数据得到的混淆矩阵。

Table 2. Confusion matrix of MLP model on test set
表 2. MLP 模型在测试集上的混淆矩阵

真值	预测值	
	低国家认同	高国家认同
低国家认同	1103	291
高国家认同	330	332

该模型整体准确率高于随机水平，κ = 29.7%。该模型的特异性(即对高国家认同个体分类的准确性)为 50.2%，而该模型的敏感性(即对低国家认同个体分类的准确性)为 81.1%，这表明如果该模型预测一名被试为低国家认同，那么其预测有 77.0%的概率是准确的。本模型的精确率(Precision)为 77.0%，本模型的 F1 分数为 78.0%。本模型 ROC 曲线下的面积即 AUC = 65%，AUC 的计算方法同时考虑了模型对于高国家认同个体和低国家认同个体的分类能力，可对模型做出合理的评价。

3.2. 国家认同的预测因子

有多种方法可以识别出机器学习模型中价值最高的预测因子，例如 LIME、iml 和 DALEX。所有的解决方法提供的都是近似解，均给出了一种可行的解决方案，但并非最优解，因此每种方法给出的预测因子的排序可能相同也可能不同。本研究使用 python 中的 DALEX 包来识别 MLP 中价值最高的预测因子。DALEX 模块原理为，每次改变一个预测因子，并评估这种改变对 MLP 模型在特定数据集上损失(loss)的影响，如果改变一个预测因子会使损失值增加更多，则认为其具有更高的价值。模型基于训练集构建，因此该分析过程在训练集上进行。表 3 列出了 DALEX 包计算出具有最高价值的 10 个预测因子。

Table 3. Based on MLP model, 10 prediction factors with the highest value for national identity are obtained
表 3. 基于 MLP 模型得到对国家认同具有最高价值的 10 个预测因子

序号	题项	Δ Dropout loss
1	家庭总收入(清理后终版)	0.5045
2	生活消费总支出(清理后终版)	0.4992
3	政府提供的经济适用房、公租房、廉租房等基本住房保障	0.4938
4	您上网进行下列活动的频率是：聊天交友(比如：微信等交友活动)	0.4937
5	社会保障是政府的基本责任，不应当由普通百姓负担(3)	0.4937
6	您认为一个好的社会应该包括哪些特征？：法治	0.4937
7	就您和您的家人来说，可能需要别人或组织提供的下列哪些方面的志愿服务？扶助残障	0.4937
8	现在的社会保障水平太低，起不到保障的作用(2)	0.4937
9	你觉得当前社会生活中以下方面的公平程度如何？：司法与执法(3)	0.4937
10	就您的个人观念来看，您能否接纳以下群体：有不同宗教信仰者(3)	0.4936

注：题项后括号中的值代表将类别变量转换为哑变量后，哑变量对应的选项。Δ Dropout loss 指如果该行提到的预测因子被置换，模型的损失值的变化。

4. 分析与讨论

在经济全球化的背景下，世界各国政治、文化、价值观相互影响，数字媒介借助信息化浪潮裹挟着不同政治取向、意识形态的海量资讯，当下民众获取信息的管道更加多元，舆论场中社交媒体大 V、意见领袖与自媒体人的非主流政治表达，冲击着公众对社会热点问题的态度与认知，也对我国国民的国家认同产生不可忽视的影响。因此，在各种思潮相互激荡的当下，把握国家认同的预测因素对妥善处理舆情、坚持主流国家认同意识的主导地位、弘扬与培育爱国主义情操具有积极意义。

已有研究者在建立回归模型的基础上利用 Shapley 值分解判断各因素的具体贡献和相对重要程度，例如基于 CSS2013 数据分析表明：对青年一代(18~35 岁)国家认同感的解释贡献率中，人口特征、地区、文化因素、经济因素和社会结构因素占比分别为 16.64%、2.71%、36.21%、29.49%、14.95%，对老一代(36 岁以上)国家认同感的解释贡献率中，人口特征、地区、文化因素、经济因素和社会结构因素占比分

别为 17.13%、8.48%、13.34%、11.53%、49.52%。也有研究者通过 Adjusted R^2 来判断自变量重要性，基于亚洲民主动态调查，有研究者指出，影响国家认同的首要因素是政治绩效。前人确定的不同模型对影响国家认同的各因素的重要性排序略有不同，这可能是由于构建模型所采用的被试数据来自不同的数据集、不同的年龄段以及纳入各模型的因素并不完全相同，但是总体上 MLP 模型确定的国家认同的预测因子与现有的研究结论是一致的。

价值最高的两个因子是“家庭总收入”与“生活消费总支出”，其代表经济因素。现有研究表明，经济因素会影响国家认同感，具体而言，收入越高国家认同感越弱，这可能是由于高收入个体获取信息的渠道更加多元，更富有批判精神，对国家认同的要求更加苛刻[5]。

第 3、5、7、8 个预测因子“政府提供的经济适用房、公租房、廉租房等基本住房保障”、“社会保障是政府的基本责任，不应当由普通百姓负担”、“扶助残障”、“现在的社会保障水平太低，起不到保障的作用”可提炼为“社会保障感知”，现有研究表明，社会保障财政支出对推动公众满意度增长具有显著效应[6]，并且，养老保障、医疗保障有助于增强政治认同和国家认同，研究显示，2006 至 2015 年社会保障公平性的提升得到了公众认可。当下我国正处于中国特色社会主义新时代，社会保障的总目标正从保障人民基本生活的兜底作用转变为增强民众向心力与国家认同的制度优势，此时党和政府再次强调共同富裕的重要性，强化社会保障，有利于增强国家认同与国家软实力。

第 4 个预测因子“您上网进行下列活动的频率：聊天交友(比如：微信等交友活动)”，这是一个双管问题，既可反映被试的网络使用频率也可以反映被试的网络社交频率，现有研究表明，国家认同与支持性社交网络呈正相关，而互联网等新兴媒介的使用会削弱个体的国家认同[7]，说明加强互联网管理与引导具有必要性。

第 6 个预测因子“您认为一个好的社会应该包括哪些特征？法治”和第 9 个预测因子“你觉得当前社会生活中以下方面的公平程度如何？司法与执法”可以提炼为“对法治的感受”。目前，有学者已经论述了“法治”和“国家认同”之间的关系，例如：有学者提出通过全面依法治国来提升国家认同[8]，亦有学者提出以法治文化建设提升公民国家认同。本研究发现：第 9 个预测因子中回答“非常公平”的被试中有 53.22%具有高国家认同，选择其他回答的被试中有 26.63%具有高国家认同。而该预测因子与目标变量之间为弱相关，相关系数 r = 0.091，基于多元线性回归模型或 Logistic 回归模型都不会将该题项作为有价值的预测因子，而机器学习模型可以捕获非线性关系。尽管存在着复杂的交互作用，但是“对法治的感受”和国家认同之间仍存在着直接关联，认同当前社会生活中司法与执法公平性的被试具有更高的国家认同，本研究提示了“对法治的感受”的重要性，与前人的回归分析结果一致，司法公平感正向预测国家认同。本研究为公众法治感知影响国家认同提供了新的证据，我国应全面推进依法治国，加快建设社会主义法治国家。

第 10 个预测因子“就您的个人观念来看，您能否接纳以下群体：有不同宗教信仰者”，该问题代表宗教因素，大量研究表明，宗教信仰与国家认同之间具有复杂的关系，21 世纪宗教对于国家与民族特性的重要性在增强。有关部门应做好对宗教的引导与规范，促进其发挥多民族共同体建构中国家的社会基础作用，团结信教群众，筑牢国家认同主流意识形态，确保国家认同统领宗教认同等其他认同。

本研究使用机器学习建模的方法补充了前人对于国家认同的影响因素的研究，具有方法学意义。目前，世界正经历百年未有之大变局，凝聚共识，提升国家认同尤为重要。在新旧矛盾相互交织的今天，传统与非传统安全问题引发关注，提升国民国家认同是维护国家安全的重要途径，而国家认同建设离不开民众参与，推进公民教育作为培育国家认同的重要手段。将国家认同的培养融入学校课程标准与教育实践，社会上广泛普及我国社会保障的基本政策以及国家在保障人民就业、住房、医疗、教育等各项工作中取得的进步。各级学校作为引导学生国家认同的桥头堡，加强爱国教育等相关德育课程的育人功能，

树立法治信念，以法治精神引领学生成才之路。进一步加强法治宣传，宣讲法治建设成果，让人民群众切身体会到社会主义法治进步，感受到更多法治温暖。澄清网络使用环境，打造清朗的网络空间，发挥社交网站的正向舆论引导功能，努力使社交媒体成为凝聚人心，增强国家认同的重要阵地，巩固国家认同主流话语在网络空间的主导地位。

5. 不足与展望

本研究使用 CSS2019 数据集，展开横截面建模分析，无法分析不同影响因素重要性随时间的变化趋势，未来可使用 CSS2013、CSS2015、CSS2017 等数据集研究中国人国家认同的影响因素随时间的变化情况。以后可考虑使用世界价值观调查(WVS)探索世界上不同国家的国家认同的影响因素之间的差异情况，展开跨文化研究。

参考文献

[1] 费孝通. 中华民族的多元一体格局[J]. 北京大学学报(哲学社会科学版), 1989(4): 3-21.

[2] 管健, 郭倩琳. 国家认同概念边界与结构维度的心理学路径[J]. 西南民族大学学报(人文社科版), 2019, 40(3): 214-221.

[3] 王玉龙, 陈阿翩, 陈慧玲. 青少年国家认同与自尊的交叉滞后分析[J]. 中国临床心理学杂志, 2021, 29(1): 148-151.

[4] 温忠麟, 黄彬彬, 汤丹丹. 问卷数据建模前传[J]. 心理科学, 2018, 41(1): 204-210.

[5] 李春玲, 刘森林. 国家认同的影响因素及其代际特征差异——基于 2013 年中国社会状况调查数据[J]. 中国社会科学, 2018(4): 132-150.

[6] 李胜会, 熊璨. 地方政府社会保障财政支出效率与满意度研究[J]. 中国行政管理, 2016(2): 104-111.

[7] 冯帅帅. 我国国民的国家认同及其影响因素研究——基于 2013 年中国社会状况综合调查数据的实证分析[D]: [硕士学位论文]. 武汉: 武汉大学, 2018.

[8] 李莉莉. 全面依法治国 提升国家认同[J]. 法制与社会, 2015(34): 137-138.