Hans Journal of Data Mining 数据挖掘, 2023, 13(2), 99-106
Published Online April 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.132010

Hans汉斯

基于BP神经网络的吉林省中学在校生规模预测研究

李晓筝

长春师范大学地理科学学院，吉林 长春

收稿日期：2023年2月27日；录用日期：2023年3月27日；发布日期：2023年4月6日

摘　要

本研究基于2005~2019年吉林省统计年鉴中的社会经济指标，利用Bootstrap样本扩充技术对数据进行数据增广，尝试应用神经网络算法构建社会经济指标与中学在校生规模关系间的反演模型，并与多因素线性回归方法建立的模型精度进行比较，意图探究社会经济基本情况对中学在校生规模的影响并针对构建的预测模型进行验证。结果发现：1) 第三产业从业规模与在校生规模的相关性优于第一产业和第二产业。2) 预测模型中的因子数 ≥ 3时，任意三因子组合的预测精度 ≥ 0.94，认为三因素模型满足预测需求为最优状态。3) 基于神经网络的科学研究技术服务及地质勘查业、交通运输仓储及邮政业和教育从业规模对中学在校生规模的预测效果最好R^2 = 0.94，RMSE = 3.98。最后根据研究发现的问题讨论形成的机制并提出具有针对性的建议。

关键词

在校生规模，神经网络，社会经济指标

Prediction Research on the Scale of Middle School Students in Jilin Province Based on BP Neural Network

Xiaozheng Li

School of Geographical Sciences, Changchun Normal University, Changchun Jilin

Received: Feb. 27th, 2023; accepted: Mar. 27th, 2023; published: Apr. 6th, 2023

Abstract

Based on the social and economic indicators in Jilin Statistical Yearbook from 2005 to 2019, the

文章引用: 李晓筝. 基于 BP 神经网络的吉林省中学在校生规模预测研究[J]. 数据挖掘, 2023, 13(2): 99-106.
DOI: 10.12677/hjdm.2023.132010

Bootstrap sample expansion technique was used to augment the data, this study attempted to use neural network algorithm to construct the inversion model of the relationship between social and economic indicators and the size of middle school students, and compared the accuracy of the model with the multi-factor linear regression method. The purpose is to explore the influence of social and economic conditions on the size of middle school students and verify the prediction model. The results show that: 1) The correlation between the employment scale in the tertiary industry and the student scale is better than that in the primary industry and the secondary industry. 2) When the number of factors in the prediction model is greater than or equal to 3, the prediction accuracy of any three-factor combination is greater than or equal to 0.94, and it is considered that the three-factor model satisfies the prediction demand as the optimal state. 3) The scale of scientific research and technical service based on neural network, geological survey, transportation and storage, postal and educational employment has the best prediction effect on the scale of middle school students, R^2 = 0.94, RMSE = 3.98. Finally, according to the problems found in the research, the mechanism of formation is discussed and the corresponding suggestions are put forward.

Keywords

Student Scale, Neural Network, Socio-Economic Indicators

1. 引言

我国教育已经进入普及化和高质量发展阶段，这不仅满足了人民群众接受更高层次教育的愿望需求，也为各行各业培养初、中、高级专业人才，成为科技进步和经济增长的重要动力。[1]在校生规模常被用来表示教育事业发展的程度；而其作为统计资料，数据的层次结构简单，在读人数只能表现数量上增减量的多少，不能明确其产生变化的原因及与社会经济因素的内部关系。社会经济和科技的发展也对教育发展产生重要的影响，社会资源配置可优化教育分配；但社会经济指标数量众多且关系复杂，而筛选出影响教育发展程度的最优指标与衡量其对教育的促进量至关重要。因此，客观刻画教育发展与社会经济的相互关系对推动教育充分、均衡发展，促进经济社会全面、协调、可持续发展具有重要的理论与现实意义。

目前关于教育规模预测方法研究还比较小众，研究的内容多围绕教育投资供给规模以提出改进措施为主，多是研究两者间的关系，而对于社会因素对教育规模的影响机制探讨较少[2] [3]。刘叔才等[4] [5] [6]利用在校生的数据资料，采用神经网络模型拟合预测教育发展规模，并与线性回归模型、回归自回归混合模型的预测结果进行了比较，结果证明基于神经网络的预测模型优于传统回归模型，但是没有关于神经网络的改进与数据筛选的研究。晏富宗等[7]在江西省在校生规模研究中也比较了传统方法与神经网络的优劣，解释了部分社会经济因素对高等教育规模的作用；王宪莲、安凤平[8]基于权重初始化－多层卷积神经网络滑动窗口融合对办学规模进行预测，提出对神经网络的改进以增加预测的准确性；但是这些研究没有涉及社会经济数据的增广问题。本文从教育规模的维度出发，使用数据增广的方案对社会经济数据进行补充，并对教育规模发展与社会经济指标展开实证检验，提出政策建议；证明基于神经网络算法构建的中学在校生预测规模能够较好地反映社会经济指标与教育规模的联系，神经网络算法可用来预测教育规模。

2. 研究设计

2.1. 实验数据和方法

本研究数据来自于吉林省统计年鉴，整理汇总了吉林省 2005 年和 2019 年的相关教育指标和社会基本情况指标。选择普通中学在校生数作为中学在校生规模以研究吉林省各区市的教育规模；从统计年鉴各种社会指标中选取了 30 个数据量充足的指标作为衡量社会发展的指标；使用 Bootstrap 样本扩充技术对数据进行增广，Bootstrap 技术依照原始样本信息进行样本扩充，可以使实验结果更稳定；[9]利用 BP 神经网络进行回归预测，BP 神经网络是一种多层前馈神经网络，对非线性系统有极强的拟合能力。

2.2. 社会指标与中学在校生规模相关性分析

社会基本情况指标与中学在校生规模的相关性分析结果见图 1，其中 80%以上的社会情况指标与中学在校生规模的相关性达到了 0.01 极显著检验水平。中学在校生规模与年末单位从业规模的相关性系数为 0.84，与交通运输仓储及邮政业从业规模、批发和零售业从业规模与金融业从业规模的相关系数为 0.79，科学研究技术服务和地质勘察业从业规模相关系数为 0.75，与教育行业从业规模的相关系数为 0.81，与卫生社会保障和社会福利业从业规模的相关系数为 0.76，与公共管理和社会组织从业规模的相关系数为 0.83，与年末邮电局数的相关系数为 0.81，与水利环境和公共设施管理也从业规模的相关系数为 0.73。

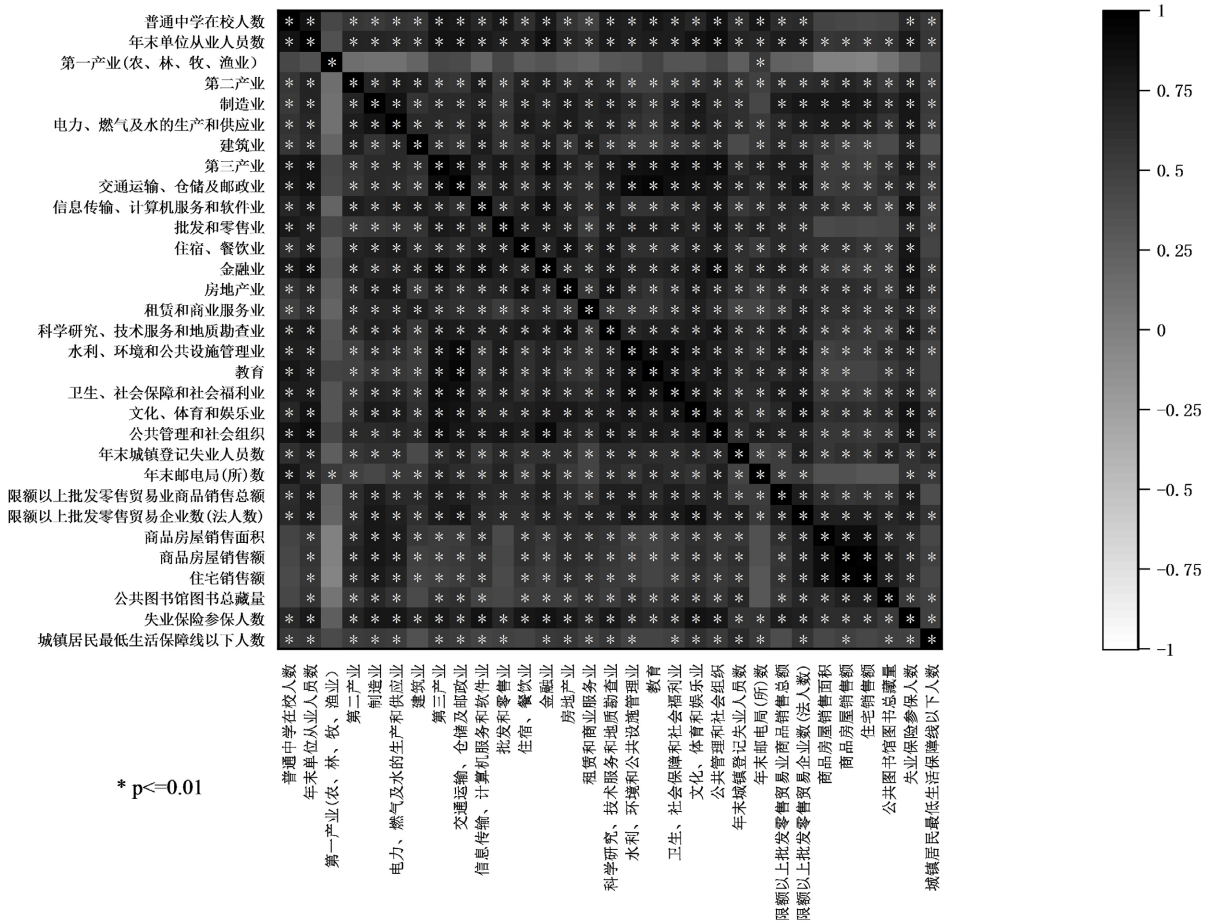

Figure 1. Social basic situation index and middle school student size correlation coefficient
图 1. 社会基本情况指标和中学在校生规模相关系数图

2.3. 基于 BP 神经网络的不同指标组合反演精度

为分析吉林省中学在校生规模与社会发展之间的关系，根据图 1 选择相关系数大于 0.75 的 11 个社会基本情况指标作为预测中学在校生规模的影响指标。对吉林省的原始样本进行增广，11 个影响指标进行随机组合，将 2047 种不同组合的影响指标分别与中学在校生规模建立 BP 神经网络预测模型。通过比较决定系数(R^2)对反演精度进行评估，结果见图 2。当影响指标的个数小于等于三个其与模型的反演精度呈正相关，精度提升的明显；当影响指标为 3 个和 4 个时，各组合平均 $R^2 = 0.96$，影响指标个数为 5 个、6 个时，决定系数 $R^2 = 0.97$，当影响指标数量为 8 个及以上时，决定系数 $R^2 = 0.98$；也就是说，当影响指标的个数大于三个，模型的预测精度已达到较高水平。再次加入指标到模型中对预测精度的提升不明显，计算量也会变得更复杂，其因增加而引起的精度波动不便于解释；因此，最佳指标个数定为三个效益最好。

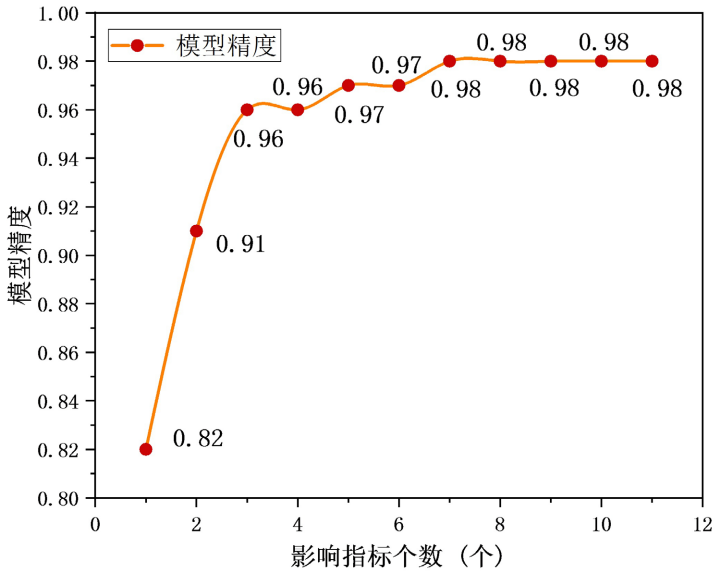

Figure 2. Influence the relationship between the number of indicators and model accuracy
图 2. 影响指标个数与模型精度关系

2.4. 神经网络预测模型构建

在确定最佳指标数的基础上，将筛选的 11 个相关性最高的数据进行单因素、两因素、三因素随机组合构建神经网络方案，结果如下：当影响指标为教育从业规模时，$R^2 = 0.939$；当影响指标为教育从业规模和交通运输仓储及邮政业时，$R^2 = 0.972$；当影响指标为交通运输仓储及邮政业、教育和科学研究技术服务及地质勘查业从业规模时，$R^2 = 0.980$。综上，本研究选择交通运输仓储及邮政业、教育和科学研究技术服务及地质勘查业从业规模三个社会情况指标设计了 3 个 BP 神经网络构建方案。

方案一：教育从业规模作为输入变量，中学在校生规模真实数值为输出变量。

方案二：教育从业规模和交通运输仓储及邮政业从业规模作为输入变量，中学在校生真实数值为输出变量。

方案三：交通运输仓储及邮政业、教育和科学研究技术服务及地质勘查业从业规模作为输入变量，中学在校生真实数值为输出变量。

3. 反演模型与结果

3.1. 神经网络反演模型验证

参照上节筛选出的三个指标，选取 2019 年相关社会统计数据对三种模型方案进行验证，通过比较中学在校生规模估计值与真实值的均方根误差(RMSE)、平均绝对误差(MAE)和决定系数(R^2)来衡量模型的精度；当 R^2 值越大、RMSE 和 MAE 值越小，模型精度越高，预测效果越好。

建模集中 3 种方案的精度差异不大，方案二和方案三的 R^2 的最高，为 0.98，方案三的 RMSE 和 MAE 均为最小值，分别为 0.36，0.30；方案一的 R^2 最低为 0.95，RMSE 和 MAE 分别为 0.77，0.66，均为建模集中最大值。综上所述，建模集中和预测集中，三者相比较而言，方案三的各项性能优于方案一和方案二(表 1)。

Table 1. Accuracy evaluation of inversion results of different schemes
表 1. 不同方案反演结果的精度评价

模型	建模集			预测集		
	R^2	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**
方案一	0.95	0.77	0.66	0.91	4.89	4.06
方案二	0.98	0.57	0.42	0.91	4.63	4.30
方案三	0.98	0.36	0.30	0.94	3.98	3.65

图 3 为中学在校生规模反演模型中实测值与预测值的散点图。预测模型的实测值样点和预测值的样点基本分布在 1:1 线附近，说明模型预测效果较好。方案三模型 R^2 为 0.94 最高，预测效果最好；方案一和方案二的 R^2 为 0.91。综上所述，方案三对中学在校生规模的预测能力较强，可以作为中学在校生规模的预测模型。

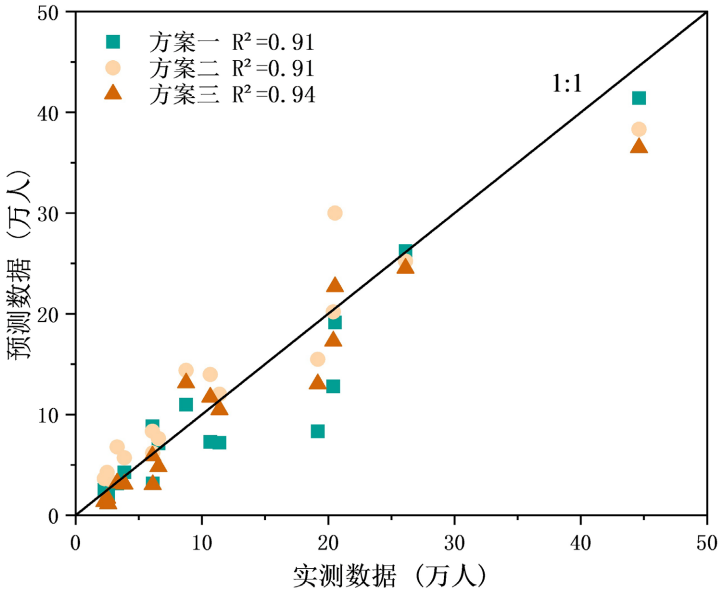

Figure 3. Scatter distribution of measured and predicted values of student scale
图 3. 在校生规模实测数值与预测数值的散点分布图

3.2. 多元线性回归反演模型与 BP 神经网络模型对比

参照表 2，选择方案三的特征参数为模型的解释变量，分别为交通运输仓储及邮政业从业规模(X_1)、科学研究技术服务及地质勘查业从业规模(X_2)和教育从业规模(X_3)，中学在校生规模为模型的被解释变量，建立 BP 神经网络预测模型和多元线性回归模型。两种模型均通过了 0.01 的 F 检验，$R^2 > 0.90$，预测模型对自变量的解释程度达到了 90%以上，其中多元线性回归模型预测集 $R^2 = 0.86$、RMSE = 4.2；BP 神经网络回归模型预测集 $R^2 = 0.94$、RMSE = 3.98；两种模型都能够很好的预测中学在校生规模。

Table 2. Prediction results of secondary school enrollment size
表 2. 中学在校生规模的预测结果

模型	回归方程	建模集		预测集	
		R^2	RMSE	R^2	RMSE
MLR	$Y = -3.940X_1 - 5.005X_2 + 5.968X_3 - 0.272$	0.94	2.27	0.86	4.20
ANN		0.98	0.36	0.94	3.98

图 4 是 ANN 预测模型和 MLR 预测模型预测值与实测值的散点图。预测值和实测值分布在 1:1 线附近，说明 ANN 模型与 MLR 预测模型具有较好的预测效果；且 ANN 比 MLR 模型预测准确度要高，稳定性更好。以上结果说明，由科学研究技术服务及地质勘查业、交通运输仓储及邮政业和教育从业规模构建的 BP 神经网络回归模型的预测能力较强。

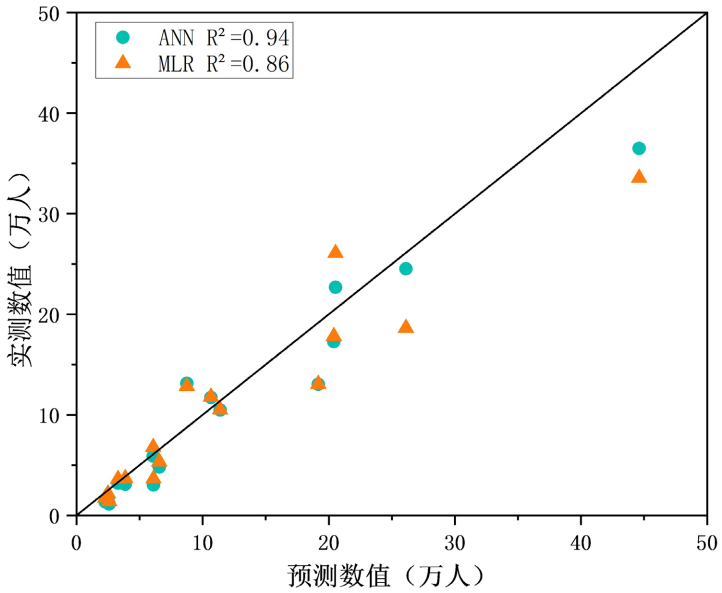

Figure 4. Comparison between neural network model and multiple linear regression model
图 4. 神经网络模型与多元线性回归模型的比较

4. 结论与讨论

4.1. 研究结论

本研究收集整理吉林省中学在校生规模和社会基本情况指标数据，通过分析中学在校生规模与社会

基本情况指标的相关性，并采用 BP 神经网络研究多种社会情况指标组合对中学在校生规模的影响，得到了对中学在校生规模影响较大的几种组合情况，最后采用 BP 神经网络预测模型和多元线性回归模型对中学规模进行对比，得出以下结论：

1) 与中学在校生规模相关性较好的社会基本情况指标分别是年末单位从业规模、交通运输类、公共管理业、金融类、科研类、教育业、社保业、水利设施管理业、文体娱乐的从业规模和邮电局数。三产与在校生规模的相关性优于一产和二产。[10]邮电局数与中学规模相关性较优。经济是教育发展的物质基础，教育机构培养劳动力的规模、层次和结构要受制于社会生产力发展水平。

2) 中学在校生规模的预测模型以科研服务及地质勘查业、交通运输仓储及邮政业和教育从业规模建立的 BP 神经网络预测模型效果较好，模型的 $R^2 = 0.94$，RMSE = 3.98，MAE = 3.65。可以准确地预测研究区中学在校生规模。学校办学规模的设置必须依存教育从业规模与学校自身的办学条件，教育从业者包括教师、教培人员以及学校其他工作人员，与学校教学规模直接相关，因此对中学的在校生规模非常敏感。而交通运输业的发展在一定程度上解决了学生上学远、上学难的问题，科学研究和技术服务业的发展使得高新技术得以在教育行业投入使用，提高教学效率，扩大教育规模，因此交通运输业从业规模和科学研究业从业规模也可以在一定程度上对在校生规模进行预测。

4.2. 讨论

值得注意的是，教育规模与社会指标的关系也反映了一系列的社会问题。教育地位随生产力发展快速提升，家庭的教育成本支出也随着学生学习压力和在校教师流失而加剧，资本无序涌入教育行业会使教育机会均衡的情况被打破，不利于社会公平。交通条件的改善，让跨区域求学机会更便捷，一面有利于解决可达性的问题，另一面安全舒适的通勤条件也可以大大提升人民的幸福感。21 世纪属于科技的时代，以交互式、多元化的高新技术对教育行业的影响尤为突出，这不仅提供了更多信息获取的途径、丰富教学内容、提升学生思维能力，且对于科技应用、知识创新等学生的全面发展有很好的促进作用。

要实现经济增长与教育发展的相互促进作用，首先必须进行制度上的创新，包括经济体制创新、基本制度创新和具体教育制度创新；其次，教育规模应适应地方经济建设和发展的需要，适应社会经济水平；其三，要贯彻落实“双减”政策，回归素质教育本身，同时应发挥教师主导教学并提高公办教师薪资待遇；最后，增加学校基础设备的投入途径，加快当地数字化硬件设施的建设，同时引导经费、技术等适当向数字化教育基础设施建设倾斜。

吉林省已出现人口低生育意向和少子化趋势，人口外流趋向明显，有户无人情况较多，在吉人口总数下滑显著；在这种人口与经济下行的大环境下，在社会因素中体现出的关于教育的一系列问题，除学费本身外，学生接受教育过程中产生的交通、补习以及数字化教育的设备费用、不均衡的地区发展等亦深刻地影响生育及人口迁移行为；[11] [12]其负向作用将导致适龄人口生育意愿下降。因此，须加大教育系统投资，大力发展公立教育，严格管理民办教育，增加学位配给制度，合理配置公立教育资源等公共服务。人口减少对教育规模的区域性差异影响较大，因统计年鉴数据缺失，本次实验未考虑适龄人口生育意愿和人口迁移等因素的影响，建议统计局增加对人口数量变动因素的统计，在后续的研究中可以考虑加入适龄人口生育意愿等因素，提高反演模型的精度。

参考文献

[1] 李晓筝. 民国时期初级中学地理教学理念的演变[J]. 地理教育, 2022(S1): 4-6.

[2] 郎益夫, 戴天虹. 基于神经网络的中国高等教育投资供给规模预测[J]. 哈尔滨工程大学学报, 2006, 27(4): 625-628.

[3] 于文波. 基于 BP 神经网络的辽宁高等教育投资规模预测[J]. 长春大学学报, 2009, 19(2): 20-22.

[4] 刘叔才, 尹平. 基于 BP 神经网络的研究生教育发展规模预测[J]. 中国社会医学杂志, 2008, 25(3): 132-134.

[5] 卢锐. 湖北省研究生教育规模预测的实证研究[D]: [硕士学位论文]. 武汉: 中南民族大学, 2011.

[6] 杨岑. 辽宁省研究生教育规模预测研究[D]: [硕士学位论文]. 沈阳: 东北大学, 2013.

[7] 晏富宗, 胡海青. 基于 BP 神经网络的区域高等教育规模预测研究——以江西省为例[J]. 教育学术月刊, 2013(12): 52-55+61. https://doi.org/10.16477/j.cnki.issn1674-2311.2013.12.018

[8] 王宪莲, 安凤平. 基于权重初始化-多层卷积神经网络滑动窗口融合的高等教育办学规模预测算法[J]. 信息技术与信息化, 2019(10): 24-29.

[9] 耿立艳, 张占福. 基于 RBF 神经网络的高等教育规模预测[J]. 科教导刊(上旬刊), 2013(3): 38+73. https://doi.org/10.16400/j.cnki.kjdks.2013.02.084

[10] 巩海霞. 教育投入的经济效应研究——基于江苏省的实证分析[D]: [博士学位论文]. 徐州: 中国矿业大学, 2009.

[11] 郭东阳. 学龄人口变动对义务教育资源配置的影响研究[D]: [博士学位论文]. 长春: 吉林大学, 2022. https://doi.org/10.27162/d.cnki.gjlin.2022.007112

[12] 王清强, 乐传永, 刘双飞. 职业院校在校生规模与国民经济发展之间的系统耦合样态分析[J]. 职业技术教育, 2021, 42(36): 39-43.