

基于机器学习的 MOOC 僵尸课程识别方法^{*}

董 晨 范馨文 郭文忠^{**} 张 凡 江秋怡
(福州大学 数学与计算机科学学院,福建 福州 350116)

摘 要 近年来,MOOC 使得学生随时都可以获取学习资源,促进知识的普及。随着 MOOC 课程数量的急剧增加,课程质量参差不齐,僵尸课程的识别对 MOOC 课程教学质量的保证及课程的改进是十分必要的。本文提出基于机器学习的 MOOC 僵尸课程识别方法 ML-MOOC,该方法具有高效、准确、低成本的特点,对教育行政主管部门对 MOOC 课程建设资金监管、教学质量监督、课程应用情况管控以及推进 MOOC 建设有重大意义;同时也对教师课程持续改进、学生选课等方面有一定的实际意义。

关键词 MOOC;僵尸课程;机器学习;教学质量;ML-MOOC

中图分类号 G420 文献标识码 A

MOOC Zombie Course Recognition Method Based on Machine Learning

DONG Chen, FAN Xin-wen, GUO Wen-zhong, ZHANG Fan, JIANG Qiu-yi
(School of Mathematics and Computer Science, Fuzhou University, Fuzhou, 350116, China)

Abstract: In recent years, MOOC has enabled students to access learning resources at any time and promoted the popularization of knowledge. With the rapid increase in the number of MOOC courses and the uneven quality of courses, the identification of zombie courses is necessary to ensure the teaching quality of MOOC courses and further improve such courses. ML-MOOC, a recognition method of MOOC zombie course based on machine learning, is proposed in this paper. This method has the characteristics of high efficiency, accuracy and low cost. It is of great significance to the education administrative department to supervise the construction of MOOC course funds, teaching quality, course application, and promote the construction of MOOC; At the same time, it also has certain practical significance for teachers to continuously improve courses and for students to choose courses.

Key words: MOOC; zombie course; machine learning; teaching quality; ML-MOOC

^{*} 收稿日期 2020-07-23

^{*} 资助项目 教育部产学合作协同育人项目(项目编号:201802076018、201802076026、201802282093);福建省一流本科专业建设项目;福建省高校在线精品课建设项目“网络空间安全概论”;福建省本科高校教育教学改革工程项目“一流课程建设‘双万计划’驱动下的《网络空间安全概论》建设”;福州大学高等教育教学改革工程项目“《网络空间安全概论》中国大学 MOOC 课程建设”;福州大学一流专业建设重点项目“‘新工科’背景下数据科学与大数据技术专业建设”;福州大学教育教学重点难点重点项目“计算机类专业产学合作综合教育实践平台建设”;福州大学一流课程建设项目“《网络空间安全概论》翻转教学模式研究”;新工科背景下网络空间安全人才培养模式的探索与实践(项目编号:E-AQGABQ20202712);国家精品在线课程“网络空间安全概论”(项目编号:0360-53007062)。

^{**} 作者简介 董晨(1979-)女,陕西西安人,讲师,博士,主要从事高等教育人才培养与专业建设研究。
^{**} 通讯作者 郭文忠

近年来,随着我国各行各业的快速发展,人们在提高生活水平的同时,对教育需求也越来越高,使得 MOOC(Massive Open Online Courses,大型开放式网络课程)这一新兴教育形式也得以发展。MOOC 是许多组织及高校免费为全世界学生提供的优质在线课程。在过去,人们若想要学习一门课程或技艺,可能通过以下几种方式:请教他人、参考相关的书籍或到学校进行学习,而 MOOC 则在这个信息泛滥的社会中为人们的学习提供了便利,将便捷可得的学习机会、灵活的学习方式与时间、低廉的学习成本、丰富优质的学习资源提供给了越来越多的学习者,为全民终身教育提供了条件和可能。如图 1 所示为 MOOC 的课程形式。各大高校进行教学的主要途径开始不局限于面对面的授课,而向着借助网络进行教学活动的方向拓展。为了借助 MOOC 优势来提高教学质量,国内众多高校开始积极探索 MOOC 的教育教学改革方法。

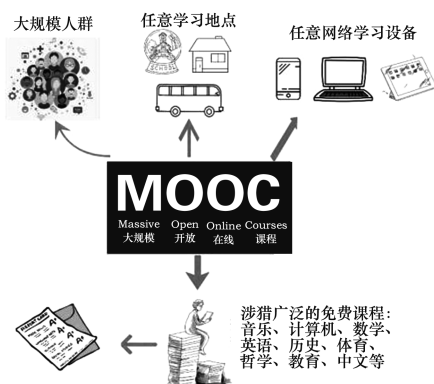


图 1 MOOC 课程形式

一、研究背景

(一) MOOC 课程建设现状

基于 MOOC 发展的广阔前景以及其为教育界带来的正面影响,我国各级教育行政主管部门大力扶持 MOOC 建设,投入大量经费予以支持。我国众多高校陆续开展了相关的网络课程建设工作,加大力度推行网络授课的教学方式。2003 年以来,我国 MOOC 形成了 3 个层次的精品课程级别体系:国家级、省级以及校级,开发了大量优质

的 MOOC 网络课程,促进了优质课程资源的传播,不断为提高我国高校教学质量的工程注入新的源泉。同时将教师的培训、网络课程开发与应用技术结合起来,致力于为学生提供一流的学习体验和学习效果。对于国内来说,优质教育资源相对匮乏,优势课程资源分布不均匀,而 MOOC 为国内高水平大学之间课程资源优势互补与共享以及向其他大学开放提供了绝好的机会^[1]。国内众多高校开始积极开展探索 MOOC 的教育教学改革方法,比如课程改革研究、学习理论研究和混合教学模式研究等,MOOC 得到了全国各高校的普遍重视。除此之外,国外对 MOOC 的相关研究比国内更充分,研究范围包括 MOOC 的理论基础、产生与发展、其对现存教育系统的影响以及高等教育、中学教育等领域的应用情况等;同时有对 Twitter、Facebook 等社交传媒在 MOOC 发展过程中的作用的研究;还有关于 MOOC 在跨国教育方面的研究等^[2]。可以说,在一定程度上,MOOC 创造了一种教育模范。

(二) MOOC 建设存在的问题

MOOC 成为众多高校教师用以辅助教学的重要方式,为学生的学习带来便捷,但是,目前 MOOC 的建设还存在一些问题:

(1) 未形成 MOOC 评价体系。目前针对各高校所开 MOOC 课程的评估机制过于简单,大部分课程是由教师或校方决定的,缺乏对学生群体的反馈做严谨的调查,且有时仅关注教学任务的完成,而忽略了教学质量的保证。我国 MOOC 研究的问题主要体现在对 MOOC 的课程完成率、课程学习效果评价、教学活动设计、讨论区交互以及学分认证等方面^[3],虽然已有研究,但还是存在着局限性。(2) 存在大量的僵尸课程。我国网络课程建设开发的总体情况还未达到一流水平,对未达标课程的监管不够,当下 MOOC 的各类课程中,除了有意义且内容充实的网课外,还充斥着许多僵尸课程。这类课程的选课人数较少,学生们对此类课程的学习主动性和接受程度均偏低,与授课教师的互动也有限。这类课程的产生是跟风建设所导致,一部分课程是课程建设者为完成某些指标或获取经费甚至获取某些荣誉而建,建成

后没有长期规划,缺乏长期的课程维护,比如线上答疑、课后讨论等。近几年对 MOOC 僵尸课程的识别与改进方法主要集中在教学改革方面,提出诸如改革课程考核方式、提升教师专业水平、重视学生意见等的改革建议^[4]。但应用现代科技手段来促进教学质量改进及提升的方法还未开始。

MOOC 现存的问题将会导致:(1)学生无法分辨课程优劣,盲目选课。僵尸课程所提供的资源甚至不能满足学生基本学习需求,使得学生查找相关学习资料、获取优质学习资源的难度增加,不利于学生自主学习和网络教学资源的有效利用。(2)教育行政部门无法监管。教育行政部门投入大量的资金,但是无法跟踪经费实施效果,无法监管课程的质量。部分僵尸课程的建设耗费大量的资金,却没有起到预期的良好教学效果。(3)网络资源爆炸,造成网络资源浪费。互联网的发展使得网络资源呈现爆炸式的增长,而泛滥的僵尸课程,占据了大量的网络资源,更是促进了这一趋势的发展。(4)无法促进建课教师对课程进行改进。课程需要根据学生的需求、知识的更新不断优化,只有这样,这门课才能得到长远的发展。没有合理的 MOOC 评价体系,无法对僵尸课程进行判别和监管,很容易造成建课教师懈怠于对课程的改进。

(三)“僵尸课程”概念的界定

僵尸课程是大学课堂及在线课程中反响比较强烈的一个普遍问题。这类课程只是为完成某些任务而开设,学生及教师参与度较低,缺乏专业的师资力量,课堂内容无趣乏味,学生对课程内容不感兴趣、不愿意听讲,收获不到真正的知识。这些课程的存在导致学生无法有效地学习到专业知识,浪费了宝贵的时间和精力,同时也浪费了学校的教学资源。随着 MOOC 课程的不断开设,课程数量急剧增加,但课程质量参差不齐,僵尸课程的识别对保证和提高 MOOC 课程的教学质量以及对此类课程进一步的改进是十分必要的。

二、利用人工智能提升教学质量的研究意义

在扩展教育的同时,不可忽视课程质量和教学效果。对于教师来说,教好一门课程主要有两

个方面:提高教学质量和保证学生学习效率。良好的教育能够激发学生兴趣和促进学生自主探究。如果学生的兴趣没有被激发,则意味着这种教育只是展示了知识或技能,而不能让学生有所收获。因此,需要开展新一轮 MOOC 课程改革,集中清理那些不符合个人培养标准的课程,其中包括那些因个人而设的或未达到质量标准的课程,要做到这些,找到一个好的甄别僵尸课程的方法是前提。它的意义有:

(1)提高教育行政部门的监管效率。如果能够对僵尸课程进行判别,就可以设置课程的优劣标准,采取有效的手段对课程进行监管。与此同时,提高经费投入门槛,杜绝僵尸课程,建设高质量的课程。(2)提高学生学习效率。MOOC 中的课程比传统课堂教育更加灵活,控制方便。若能使网络课程的质量得到保证,学生可直接获取优质学习资源,将自身筛选优质课程的时间节省在学习上,使学生在相同时间内获取更多的信息。同时使学生可以参与网络教材增删修改的过程,传统教学中一些僵化陈旧的部分也会因此而更新。(3)提高 MOOC 自身的竞争力。学生是所有在线教育的基础,只有不断提高平台中的课程质量,消除滥竽充数的僵尸课程,才能让 MOOC 对学生具有吸引力,让学生愿意将宝贵的学习时间用在 MOOC 平台开展的学习活动上,MOOC 这个平台才有继续存活与发展的可能。(4)有利于鼓励和督促教师认真授课、尊重课堂,使网课真正达到学生的需求,而不是迫于工作任务参与选课,或者为了满足学分要求而上课;同时也避免出现课程内容千篇一律、照搬课本内容的现象,使教学真正做到以学生为中心,加强师生之间的交流互动,调动学生的主观能动性。(5)提高优质网课的传播,有利于实现全民学习、终身教育;将打破传统授课方式的限制,有利于改变教育方式。互联网的快速发展,使在校大学生可以参与国内外知名专家、教授的授课课堂,学得专业知识,甚至可以与领域内的硕士、博士以及其他高层次人才进行交流探讨。(6)应用现代科技手段来促进教学质量改进及提升。在学术界,人工智能技术是一种融合多门学科知识的先进技术,它可以使机器胜任一些

通常情况下需要人类才可以完成的复杂工作。采用这种科学的手段,对教学质量进行提升和改进是高效可行的,对教学的创新有推动作用。

三、ML-MOOC 技术方案

(一) 方法基本原理

BP (Back Propagation) 神经网络是一个多层的前馈神经网络,它的主要特征是:信号前向传播,误差反向传播。举例来说,如果一个神经网络模型中只含有一个隐藏层,BP 神经网络的运行主要分为以下两个阶段:(1) 信号前向传播。信号将从输入层发出,经过隐藏层,到达输出层,当预期设定的输出与现实输出存在误差时,则将开始第二阶段的传播;(2) 误差反向传播。从输出层到隐藏层,最后到达输入层,过程中各层将根据误差来调整权值和阈值,使误差梯度下降。通过反复的训练,当达到最小误差标准时,即可确定最终各层间的权值和阈值,停止训练。图 2 为一个三层 BP 神经网络,其中 x_i 为输入层输入, w_{ij} 为神经元间的权值, b_i 为阈值。有监督 BP 神经网络有九步学习过程^[5]:

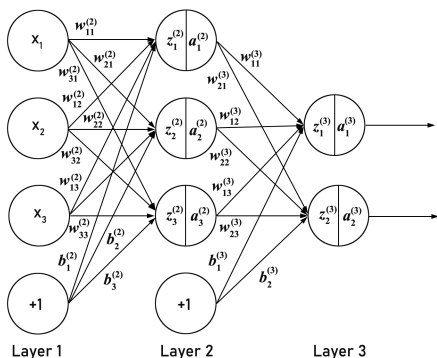


图2 三层 BP 神经网络模型

第一,神经网络的初始化。设定最大训练次数、误差函数等参数,将连接权值设置为定义在区间 $(-1,1)$ 内的随机数值。

第二,随机选取 m 个样本,对于其中每一个样本 x 及期望输出 d ,如式(1)和(2),其中 n 表示输入层节点数, q 表示隐含层节点数:

$$x = (x_1, x_2, \dots, x_n) \quad (1)$$

$$d = (d_1, d_2, \dots, d_q) \quad (2)$$

第三,得出隐含层每个节点的输入 $z_h^{(2)}$ 和输出 $a_h^{(2)}$,如式(3)和(4),其中 p 表示隐含层节点总数, h 表示当前隐含层节点:

$$z_h^{(2)} = \sum_{i=1}^n w_{ih} x_i - b_h \quad h = 1, 2, \dots, p \quad (3)$$

$$a_h^{(2)} = f(z_h^{(2)}) \quad h = 1, 2, \dots, p \quad (4)$$

第四,得出输出层每个节点的输入 $z_o^{(3)}$ 和输出 $a_o^{(3)}$,如式(5)和(6),其中 q 表示隐含层节点总数, o 表示当前隐含层节点:

$$z_o^{(3)} = \sum_{h=1}^p w_{ho} a_h^{(2)} - b_o \quad o = 1, 2, \dots, q \quad (5)$$

$$a_o^{(3)} = f(z_o^{(3)}) \quad o = 1, 2, \dots, q \quad (6)$$

第五,利用期望结果与实际结果间的误差,计算出 e 对输出层节点的偏导数 δ_o ,如式(7):

$$\delta_o = \frac{\partial e}{\partial z_o^{(3)}} = \frac{\partial [\frac{1}{2} \sum_{o=1}^q (d_o - a_o^{(3)})^2]}{\partial z_o^{(3)}} \quad (7)$$

第六,通过输出层节点的偏导数 δ_o 和隐含层节点的输出值来重置两层神经元间的权值 w_{ho} ,给定学习率 η ,如式(8)和(9):

$$\Delta w_{ho} = -\eta \frac{\partial e}{\partial w_{ho}} = \eta \delta_o a_h^{(2)} \quad (8)$$

$$w_{ho}^{N+1} = w_{ho}^N + \eta \delta_o a_h^{(2)} \quad (9)$$

第七,通过误差函数对隐含层节点的偏导数 δ_h 和输入层各节点的输入来重置该两层神经元间的权值 w_{ih} ,如式(10)和(11):

$$\Delta w_{ih} = -\eta \frac{\partial e}{\partial w_{ih}} = \eta \delta_h x_i \quad (10)$$

$$w_{ih}^{N+1} = w_{ih}^N + \eta \delta_h x_i \quad (11)$$

第八,用误差函数计算全局误差,其中 m 为样本个数,对于每一个样本,误差函数 e 如式(12):

$$e = \frac{1}{2} \sum_{o=1}^q (d - a_o^{(3)})^2 \quad (12)$$

全局误差为 m 个样本误差的平均值。

第九,若达到停止条件,即此时误差小于预设的精度,或此时训练次数达到预设的最大次数,则结束算法。否则,训练下一个学习样本。

(二) ML-MOOC 僵尸课程识别方法

基于 BP 神经网络的僵尸课程分类算法包含

以下 3 个步骤:构建 BP 神经网络模型、BP 神经网络训练以及和神经网络分类。

1. 算法流程图

首先,观察僵尸课程与国家精品课的明显区别点,从而提取僵尸课程的各种显著特征;然后设计好神经网络模型,利用所收集的数据集训练神经网络,得到特征数据与是否为僵尸课程的一一对应关系。在实际判断过程中,只需收集目标课程的特征数据,就可以从所训练好的神经网络中得出对应的结果。算法流程图见图 3。

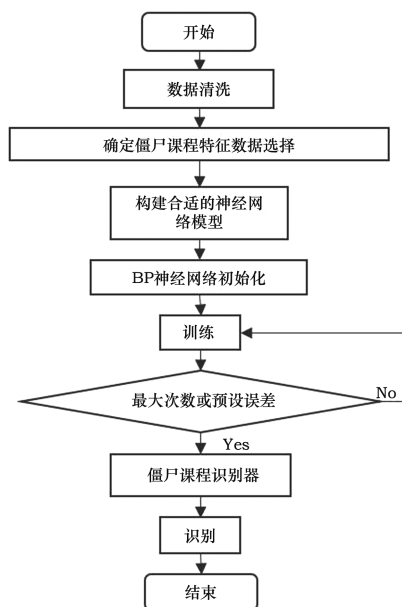


图 3 ML-MOOC 僵尸课程识别算法流程图

2. ML-MOOC 算法步骤

本文提供一种基于机器学习的 MOOC 僵尸课程识别方法,具有高效、准确、低成本的特点,对推进 MOOC 建设有重大意义。这种课程识别方法包括以下步骤:

(1)数据采集。从 MOOC 平台上采集若干样本课程。(2)数据清洗。清除重复、疑似错误、偏离整体分布的样本,使数据都是意义明确的、正确的、规范的,提高数据的可用性。(3)构造数据集。提取样本课程特征并添加标签,构造训练数据集,所述样本课程特征包括已选课人数、总评论数、总评论回复数、平均评论回复数、平均投票数和教师参与评论数。(4)构建 BP 神经网络模型。

如图 4 所示,输入每组样本数据的各项特征指标,其中输入层节点数为 6;输出该课程是否为僵尸课程的判断结果,其中输出层节点数为 1。隐含层节点数为: $l = \sqrt{n + m} + a$, 其中, n :为输入层节点数, m 为输出层节点数, a 为区间 $[1,10]$ 内的常数。所述隐含层以及输出层的激励函数采用的是 Tansig 函数。(5)训练 BP 神经网络模型。根据训练数据集中的数据初始化并训练,得到训练后的 BP 神经网络模型。(6)测试。输入待测的 MOOC 课程,判断是否为僵尸课程。

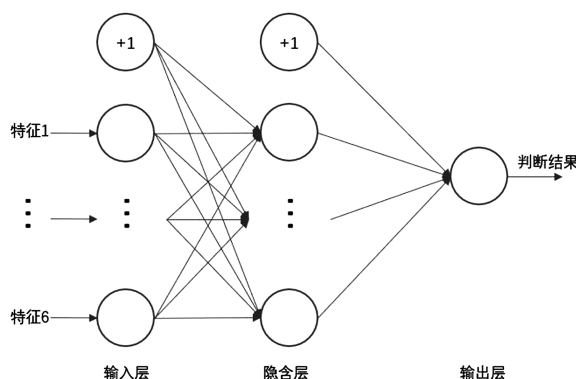


图 4 BP 神经网络示意图

四、实施样例分析

(一) 实验环境与配置

本文实验选用 Matlab 中的神经网络工具箱进行神经网络分类器的训练。在 Matlab R2017a 的环境下,使用 Matlab 代码编写算法对应程序进行仿真,处理器为 Intel(R) Core(TM) i5-8250U CPU@ 1.60GHz,操作系统为 Windows 10 系统。

(二) 训练

收集 120 门样本课程的各项特征数据,将前 110 门课程的相关数据作为输入,特征包括已选课人数、总评论数、总评论回复数、平均评论回复数、平均评论投票数、教师参与评论数,将判定各门课程是否为僵尸课程的结果作为输出。将选课人数少于 3 000 人次的课程判为可能的僵尸课程,若在此基础上,评论区总评论数少于 60 条,则将其选作为僵尸课程。若判定此门课程是僵尸课程,该输出数据为 0,反之若判定此门课程不是僵

尸课程,则输出数据为1。

将如表1所示的部分训练集作为训练数据输入初始化的BP模型,将表2所示的测试集作为测试集输入训练好的BP模型。如图5(a)所示,这是某一次测试的均方误差图,均方误差越小,说明实际输出和期望输出间的误差值越小。可以看出在训练次数到达6328时,神经网络达到最优目标误差。图5(b)是某一次测试的回归图,通过绘制回归线来判断训练数据的拟合程度。可以看出拟合程度好,预设的僵尸课程与非僵尸课程的

数据分别集中于回归线的两端点,区分度明显,训练效果显著。

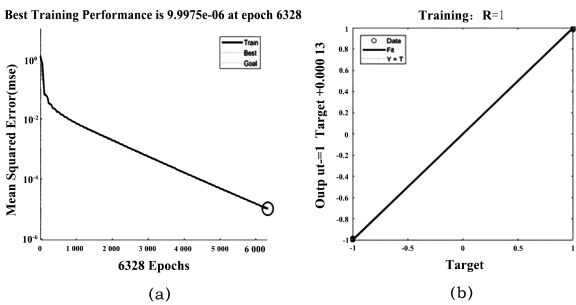


图5 测试均方误差图和回归图

表1 部分训练样本数据集

课程名称	已选课人数 /人	总评论数 /条	总评论回复数 /条	平均评论回复数 /条	平均评论投票数 /条	教师参与评论数 /条	标签
大学英语(口语)	470 525	439	6 619	15.077	1.763 0	109	1
金融学	330 430	677	1 101	1.626	0.456 0	349	1
沟通心理学	206 431	655	5 192	7.927	0.672 0	55	1
医学统计学	113 856	261	4 012	15.372	0.410 0	12	1
生命科学伦理	1 060	3	1	0.333	0.000 0	0	0
采矿学	1 227	2	2	1.000	0.000 0	2	0
营养学	108 094	775	11 690	15.084	0.974 0	259	1
计算机操作系统	55 948	69	106	1.536	0.246 0	0	1
概率论与数理统计	68 513	735	2 857	3.887	0.392 0	3	1
农业植物病理学	37	27	0	0.000	0.000 0	0	0
电磁场与微波技术实验	1 058	21	575	27.381	0.009 5	3	0
环境伦理学	624	10	2	0.500	0.100 0	5	0
绿色中国	1 130	32	599	18.719	0.313 0	0	0
饲料毒物学	885	7	3	0.429	0.000 0	1	0
动物流行病学	1 242	5	4	0.800	0.000 0	1	0
大观与微观	167 854	836	2 824	3.398	1.075 0	26	1
细读张爱玲	162 117	1 333	8 000	6.002	1.254 0	19	1
现代礼仪	100 282	836	3 660	4.378	0.289 0	13	1
微观经济学	99 085	391	442	1.130	0.402 0	6	1
全球治理与国际组织	817	4	3	0.750	0.000 0	2	0
制度与人情	1 396	8	49	6.125	0.125 0	1	0

表 2 测试样本数据集

课程名称	已选课人数 /人	总评论数 /条	总评论回复数 /条	平均评论回复数 /条	平均评论投票数 /条	教师参与评论数 /条	判定结果
现场生命急救知识技能	39 210	128	5 935	46.367	1.109	56	1
急救常识	35 332	705	1 957	16.960	0.681	2	1
人体解剖学	27 067	205	308	1.502	0.141	173	1
生理学	31 076	244	4 211	17.258	2.664	18	1
解剖与临床	30 474	193	1 561	8.088	0.845	154	1
家畜寄生虫病学	958	16	144	9.000	0.000	7	0
普通生态学	940	42	65	1.548	0.119	15	0
农村社会保障	392	89	359	4.034	0.371	0	0
XML 与电子商务应用	522	7	2	0.286	0.000	0	0
青草药识别	890	23	463	20.130	0.652	22	0

(三) 测试

对同一测试数据进行 8 次测试,非僵尸课程的期望输出为 1,急救常识、人体解剖学、解剖与临床 3 门课程 8 次测试中的实际输出均为 1,与期望输出无误差,现场生命急救知识技能课程距期望输出最大误差为 0.001,这 5 门课程中最大的误差出现在生理学课程,实际输出为 0.971 2,误差也仅为 0.028 8。可以看出,僵尸课程期望输出为 0,家畜寄生虫病学、普通生态学、XML 与电子商务应用 3 门课程 8 次测试中的实际输出均为 0,与期望输出无误差,农村社会保障、青草药识别两门课程在这 8 次测试中均出现误差,但最大的误差也仅为 0.014 7。经过多次统计测试的输出结果,基于 BP 神经网络训练的僵尸课程分类器的识别正确率接近 100%,如图 6 所示。

五、结束语

本研究结果说明在识别 MOOC 课程中的僵尸课程问题上,应用人工智能的方法是科学可行的。这种方法可以更好地约束和监管我国 MOOC 的管理工作,鉴别并筛除其中一些劣质课程,确保教师的教学效果和学生的学习质量,使得 MOOC 在我国的发展步上一个新的台阶。

参考文献:

[1]江志斌,丁洁,余建波,等. 基于 MOOCs 的教育教学改革[J].

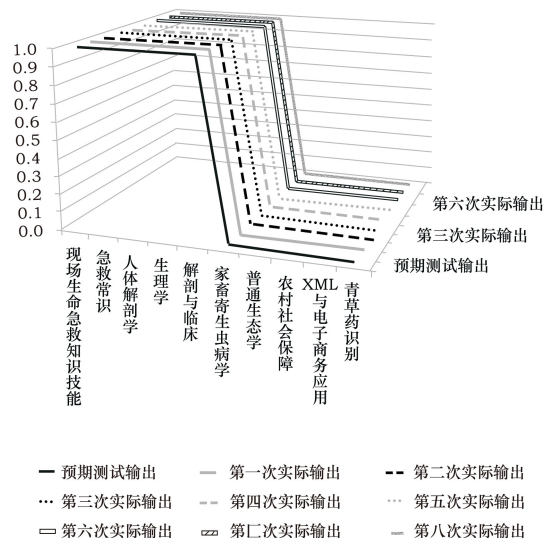


图 6 八次输出数据折线图

高等理科教育,2014,116(4):73-77.

- [2]李木子. 基于 MOOC 的网络教学模式研究[J]. 计算机教育, 2014(9):34-36.
- [3]张红艳,杨丽雪. 我国 MOOC 研究现状及存在问题的分析[J]. 开放学习研究,2017(6):35-42.
- [4]JHONY CHOON YEONG NG,陈红雪,宋凯文,等. 基于扎根理论的高校“水课”教育改革研究[J]. 高教学刊,2017(22): 15-18.
- [5]郑君里,杨行峻. 人工神经网络[M]. 北京:高等教育出版社,1992:15-30.

(责任编辑 李世萍)