

网络学习资源进化预警模型设计

杨现民, 袁 萌, 李康康

(江苏师范大学 江苏省教育信息化工程技术研究中心, 江苏 徐州 221116)

[摘 要] 大规模开放协同环境给网络学习资源的高效管理带来了新挑战, 基于浏览量与评分排名的传统管理方法难以保证高质量网络学习资源的生成与持续进化。如何建构智能化的网络学习资源进化预警模型, 是大数据时代破解网络学习资源高效管理的核心问题。该研究以网络学习资源进化预警任务的实施流程为依据, 在探讨资源进化态表征、资源进化要素项提取、资源进化态标注、资源进化预警任务的数学定义等问题的基础上, 构建了网络学习资源进化预警的技术模型。通过学习元平台的两个数据集实验验证表明, 研究提出的预警模型在预测精准率、F1 和 AUC 等指标方面表现优于对比算法, 能较好地完成网络学习资源进化预警的任务。

[关键词] 网络学习资源; 学习资源进化; 预警模型; 深度学习; 卷积神经网络

[中图分类号] G434

[文献标志码] A

[作者简介] 杨现民(1982—), 男, 河北邢台人。教授, 博士, 主要从事智慧教育、教育大数据、网络学习资源研究。E-mail: yangxianmin8888@163.com。

一、引言

新冠肺炎疫情对各级各类教育系统都产生了巨大冲击, 凸显出海量网络学习资源进化不足导致的质量不佳与个性化严重缺失的问题, 因此, 网络学习资源进化的研究变得尤为重要。当前, 虽然在学习资源进化的概念模型、关键技术以及单模态资源进化规律等方面的研究取得积极进展, 但在网络学习资源进化预警的关键技术方面尚未取得突破, 而预警机制的建设对于高效管理海量学习资源、持续提升资源品质, 具有至关重要意义。当前网络学习领域的预警技术, 主要以学习者作为预警对象, 通过传统的机器学习算法建立预警技术模型, 应用于诸如课程成绩预警、学业辍学预警等教学服务场景。由于网络学习资源进化影响因素的复杂性与动态变化特征, 当前以学习者作为预警对象的传统机器学习算法模型, 难以实现网络学习资源进化的精准预测和预警信息的有效反馈。基于此, 本研究尝试结合卷积神经网络和 Word2Vector 技术, 构建网络学习资源进化预警模型, 以期教育信

息化 2.0 时代下大规模网络学习资源的质量提升与智能管理提供技术支撑。

二、相关研究

(一) 网络学习资源进化研究现状

学习资源进化是指学习资源为满足学习者各种动态、个性化的学习需求而进行自身内容和结构完善及调整的过程^[1]。网络学习资源包括开放知识社区资源、网络课程资源、学习软件资源等多种形态, 每种资源的进化方式和特点各不相同。典型的开放知识社区资源如 Wiki 百科、学习元、百度词条等, 其进化方式是由普通用户对同一词条进行编辑修改以实现词条内容的不断完善^[2]。词条内容的潜在专业性要求促使具有相关领域知识的用户群体协同维护同一内容, 保证了资源更新的及时性、前沿性和完善性。但因用户群体的知识水平不平衡性和匿名性, 开放知识社区资源存在一定的质量问题 and 信任度问题^[3]。因此, 部分学者提出通过资源内容协同编辑和版本控制技术等方式改善词条质量^[4-5]。网络课程资源和学习软件资源的

进化方式主要由专业团队设计、制作、发布和更新,故此类型资源结构都相对比较稳定,但其进化周期较长、迭代成本较高^[9]。

近年来,围绕学习资源进化模型的研究主要从生态学视角探讨资源的内容进化和资源的关联进化。从生态学角度看,学习资源类似生命有机体,同样具有生命周期,主要呈现“产生—生长—成熟—衰退—淘汰”的一般发展轨迹^[7]。徐刘杰等^[8]以后现代知识观为指导,建构出泛在学习资源进化的动力模型,提出学习资源进化动力分为内外两种动力,分别是由用户选择和环境选择构成的资源进化外部动力以及由资源之间的竞争与协同(自组织行为)构成的资源进化内部动力。从资源进化的内容更新迭代来看,针对大规模开放环境下学习资源内容进化质量难以控制的问题,有学者提出基于语义基因和信任评估的内容进化智能控制模型^[9]、网络信息资源利用与再生的一般模型^[10]等研究,从版本控制技术和内容参与方行为意向等方面改善资源内容进化。从资源关联进化来看,针对在资源关联进化中的静态关系元数据无法有效描述语义层关系的问题,有学者提出利用语义本体^[11-12]、关联数据^[13-14]、数据挖掘^[15]、知识图谱^[16-17]等技术研究学习资源的关联网络,以增强泛在学习资源进化“健壮性”。

总体而言,当前大多数的网络学习资源进化模型属于概念性模型,缺乏进化状态的量化表征和统计数据,亟须深入开展网络学习资源进化预警的实证研究,研发网络学习资源的预警技术模型,提升大规模网络学习资源的进化质量和管理效能。

(二)网络学习中的预警技术研究

当前国内外教育领域有关预警技术的研究,主要集中于网络学习方面,相关预警技术对于网络学习资源预警模型的构建具有借鉴意义。近年来,为应对网络学习中课业完成率低等问题,相关机构已研发出一些学习预警产品,如 Course Signals、Starfish Early Alert System 等。当前网络学习预警研究主要采用数据挖掘技术来实现。研究者们通过整合学习者的背景信息、学习成绩历史数据以及课程学习的过程性数据,选择适合的数据挖掘方法,比如预测/分类、聚类、关联分析、时间序列分析等,构建学习预警模型,以实现课程成绩预警^[18]、学业表现预警^[19]与辍学预警^[20]。在具体挖掘算法的选择上,使用最多的是线性回归、逻辑回归、神经网络、决策树和支持向量机等算法。由于研究情境、面向对象以及运行平台的差异,无法确切地判定哪一种数据挖掘算法是最佳的。因此,研究者们大多综合比较多种算法来构建学习预警模型,以期达到特

定情境下的最优预警效果。例如,Kotsiantis^[21]通过比较 M5 树模型、M5 规则树模型、神经网络、线性回归、局部加权线性回归、支持向量机等六种技术,发现 M5 规则树模型的成绩预测准确率最高,且可理解性最好;Hu 等^[22]搜集了一门本科课程所有的学习活动数据,比较 C4.5、CART (Classification And Regression Tree) 和逻辑回归三种技术构建预警模型的准确性,最后发现集成了自适应增强技术的 CART 模型效果最好。

随着人工智能技术的快速发展及其在教育领域的拓展应用,采用神经网络模型开展在线学习成绩预警的研究也日渐增多。例如,李景奇等^[23]开发了基于 BKT(Bayesian Knowledge Tracing)的网络教学跟踪评价模型,可以实现网络学习的成绩预测与预警;Yang 等^[24]使用在线视频学习行为数据训练了一个时间序列神经网络,用来对 MOOC 学习者的成绩进行动态预警;陈晋音等^[25]基于在线学习行为特征数据,提出了一种基于 BP 神经网络的学习成绩预测方法。除了研究在线学习成绩预警外,有学者开始关注网络学习过程中的情绪与情感状态监测及其预警。例如,Fei 等^[26]以 MOOC 平台中的课程论坛交互文本数据为样本,融合 Word2vec 模型与机器学习算法构建了学习者情感分类器,用于预测 MOOC 学习者的情感变化;Wang 等^[27]利用语义分析模型来跟踪在线学习者的情感倾向,并通过情感量化机器学习方法构建了学业预警模型,能够及时发现无法正常完成学业的学生;还有学者^[28]结合情感分析技术,建立机器学习模型,识别在线学习者的抑郁、焦虑等心理状态,进而开展多层面的心理预警。总体来看,网络学习预警研究呈现出从认知层面到非认知层面、从经典数据挖掘算法到神经网络算法不断拓展和升级的趋势。

综上可知,当前网络学习领域的预警研究主要以学习者作为预警对象,通过多维度数据(如成绩、行为、社会特征等)的挖掘,构建课程成绩预警模型、学业表现预警模型、学生辍学预警模型以及学生心理预警模型,对于改善网络学习品质起到了积极作用。然而,从预警技术来看,仍以传统的机器学习算法为主,在前沿的深度学习算法研究与应用上仍有很大发展空间。此外,学习资源也是网络学习生态的关键要素,对资源的生成与进化进行动态预警,是解决当前教育信息化领域资源更新缓慢、应用率偏低、管理效能低下等现实问题的重要途径,是一项极具挑战性的新课题。

三、网络学习资源进化预警模型构建

为科学地构建网络学习资源进化的预警模型,本

研究设计了完整的网络学习资源进化预警流程,如图1所示。该流程主要涉及数据采集、数据处理、数据聚类、预警任务数学定义、深度学习模型、模型训练和实验、预警应用七个步骤。在预警任务执行中涉及几个基础性问题需要明确,分别是资源进化态如何表征、需要抽取哪些资源进化要素项、资源进化态如何标注以及资源进化预警任务如何进行数学定义,最后才能构建完整的资源进化预警模型。

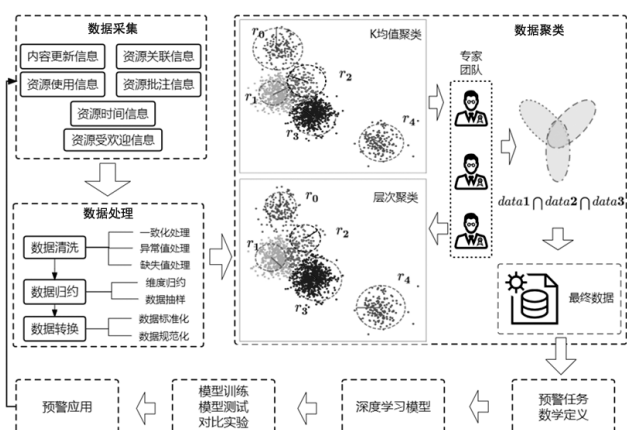


图1 网络学习资源预警任务的整体流程

(一)进化态表征

从生态学角度看,泛在学习资源生态中网络学习资源进化包括内容进化和关联进化两种模式。在进化过程中,网络学习资源进化过程类似生命有机体的发展过程一般,同样经历“初始、生长、健壮、衰老和死亡”的生命周期,呈现出“产生—生长—成熟—衰退—淘汰”的一般发展轨迹。从网络学习资源进化的动力模型看,学习资源的进化动力来源于由用户选择和环境选择构成的资源进化外部动力以及由资源之间的竞争与协同(自组织行为)构成的资源进化内部动力,这两种动力构成学习资源进化的生命力,促使学习资源不断地进化发展。学习资源在进化过程中,受两种动力此消彼长的影响,其发展轨迹相对于生命体的发展也更加复杂多变,但总体趋势不会跳出一般的发展轨迹。

基于以上认识,我们将网络学习资源进化的状态划分为:初始态、成长态、稳定态、衰退态、死亡态五种状态,分别编码为 $y_i=\{0,1,2,3,4\}$ 。如图2所示,网络学习资源进化过程大致可分为三种情况:一是学习资源自创建以来,几乎没有内容建设和更新,信息量始终处于较低的水平,如图2类别1曲线;二是学习资源自创建以来,一直处于建设、更新、被浏览和关注的状态,信息量由不断增加到处于稳定的状态,如图2类别2曲线;三是学习资源经历了成长态、衰退态和

死亡态的过程,这是因为学习资源在一段时间内关注度攀升,但随着资源建设出现滞后、停滞,出现较大纰漏等问题,使得用户点踩数、负面评论逐渐增多,信息量呈现负增长状态直至死亡态,如图2类别3曲线。归纳分析三种不同类别的网络学习资源的进化进程,总结各进化态定义如下:(1)初始态表示网络学习资源初始创建的状态,此时学习资源的各项指标尚处于初步积累阶段,由资源进化相关的数据特征信息计算所得的信息量处于起始状态;(2)成长态表示学习资源不断被用户选择、使用、更新、评论,且呈现出信息量不断增长的状态,此时学习资源的各项指标较为活跃;(3)稳定态表示学习资源被用户使用、更新的频率处于平缓稳定状态阶段,且与资源进化相关的信息量持续处于较高的活跃期;(4)衰退态表示信息量不断降低的状态阶段,此时因学习资源内容过时或结构不合理等原因,用户的负面信息或整改意见不断增加,而学习资源迟迟不进行优化,导致信息体量出现降低的现象;(5)死亡态表示学习资源基本没有用户选择、使用、更新或评论的状态阶段。死亡态包括两种情况:一种是学习资源信息体量由衰退态逐渐退化到信息量长期处于较低的信息量阶段;另一种是学习资源自创建以来,始终处于极少量用户选择、使用、更新或评论的状态阶段,表现为信息量长期处于低水平建设状态。

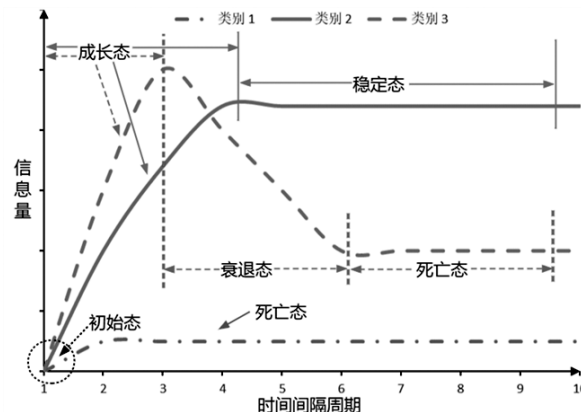


图2 网络学习资源进化态示意图

(二)进化要素项提取

网络学习资源包含种类繁多,各自媒体的表征形式与结构复杂性都不相同,为深入聚焦某一种网络学习资源进化预警技术的研发,本研究表述的网络学习资源特指开放知识学习资源,例如维基百科、学习元等平台产生的学习资源。在开放知识学习资源方面,反映此类学习资源进化状态的要素有很多,因此需要对共性要素指标进行总结分析,筛选影响学习资源进化的要素组合成数据的特征空间。通过对学习资源概

念模型、资源进化模式以及资源进化评价指标的前期研究,本研究确定从七个方面的资源进化要素项评估学习资源的进化状态。具体要素项见表1。这些资源进化要素项包括:内容更新信息、资源关联信息、资源使用信息、资源批注信息、资源评分信息、资源受欢迎信息和资源时间信息。每一项进化要素都有具体的细化项指标与之对应。因此,可依据该细化项指标采集数据,组成数据的输入空间进行特征提取和表征,通过构建深度学习模型,预测网络学习资源处于哪个阶段的进化状态,以实现衰退态或死亡态的及时预警。

(三)进化态标注

依据网络学习资源进化要素细化项指标,在收集初步的数据之后,还需要对数据进行数据清洗、归约和转换等数据处理操作。数据清洗处理主要包括一致化处理、异常值处理和缺失值处理。一致化处理目的是解决数据采集过程中数据类型不一致问题。针对缺失值,本研究主要采取补0操作。通过数据分析所得的异常数据,则采用舍弃的方法。数据归约主要是通过维度归约和数据抽样的方法,降低数据的冗余,使数据更简洁。数据转换则通过数据标准化和规范化操

作,将数据映射到较低取值范围。例如,可以将资源评分信息标准化为[0,1]区间,将资源时间信息规范为以“周”为单位的整数时间间隔等操作。

在获得统一、规范的数据之后,开始标注学习资源的进化态。本研究以学习元平台为依托,采集了开放知识社区资源的相关数据。据数据统计分析显示,目前已获得约17万的数据可作为实验基础数据。显然,完全以人工标注的方式评估学习资源的进化态是不现实的。因此我们设计用无监督学习之聚类算法+专家抽检审阅相结合的方法,对数据进行标记,以减少人工标记的工作量。具体流程如图3所示:(1)确定中心。将数据分别通过K均值聚类算法和层次聚类算法获得5类簇群和簇群中心的信息。如图3中不同颜色、形状的数据点表示经过聚类算法智能分类得到的不同类别的数据和聚类中心点。(2)设定半径。为进一步确定自动分类正确性,采取聚类中心画圆的方式进行二次筛选确定。各簇以簇群中心和半径 r (人工设定)画圆。(3)专家筛选。任意选取距离圆心由近及远的 n 条数据分别呈现给专家组,邀请专家依据各进化态的定义,判定该圆内数据的进化态。若筛选过后的数据100%属于某一进化态类别,就可将此半径为 r

表1 网络学习资源进化要素项

进化要素	进化要素细化项	要素描述	示例
内容更新信息	内容编辑次数	资源内容被编辑的次数	用户A对资源R进行了内容编辑(类似维基中的词条编辑),计数增1
	版本更新次数	资源版本更新的次数	用户A对资源R进行重新上传,计数增1
资源关联信息	关联其他知识社区数量	资源与其它资源建立关联的数量	资源R1和资源R2之间新增关联,则R1和R2的关联信息各增加1
	关联知识群数	资源与其它知识群建立关联的数量	资源R与知识K新增关联,则R的知识群数增加1
资源使用信息	资源浏览量	资源被用户浏览的次数	用户A浏览资源R一次,计数增1
	资源收藏量	资源被用户收藏的次数	用户A下载资源R一次,计数增1
资源批注信息	资源批注量	资源被成员批注的次数	成员A对资源R批注一次,计数增1
资源评分信息	资源评分数量	资源被用户评分的次数	成员A对资源R评分一次,计数增1
	准确性平均分	资源在一定时间间隔内,用户对资源准确性评分的平均分	准确性评分之和除以该时间间隔内的评分数量
	客观性平均分	资源在一定时间间隔内,用户对资源客观性评分的平均分	客观性评分之和除以该时间间隔内的评分数量
	完整性平均分	资源在一定时间间隔内,用户对资源完整性评分的平均分	完整性评分之和除以该时间间隔内的评分数量
	规范性平均分	资源在一定时间间隔内,用户对资源规范性评分的平均分	规范性评分之和除以该时间间隔内的评分数量
	及时性平均分	资源在一定时间间隔内,用户对资源及时性评分的平均分	及时性评分之和除以该时间间隔内的评分数量
资源受欢迎信息	点赞数	资源被用户点赞的次数	成员A对资源R点赞一次,计数增1
	点踩数	资源被用户点踩的次数	成员A对资源R点踩一次,计数增1
资源时间信息	资源创建时间	资源创建的时间	2021-01-01 08:00:00
	资源采样时间	要素信息采样的时间	2021-05-01 19:30:00

的圆内数据分门别类做标注,否则缩小半径值,重新选定数据。(4)设定圆环。为尽可能覆盖正确分类的范围,采取圆环不断拓宽、专家随机挑选审核的方式进行标注。例如,任意选取半径大于 r 且小于 $r+\delta$ (人工设定)圆环,随机选取圆环内距离圆心由近及远的 n 条数据分别呈现给专家组,由专家组判断该圆环内数据全部属于半径 r 圆内数据同一类别。若该圆环内数据与半径为 r 的圆内数据类别一致,继续拓宽 δ 长度的圆环,判定新圆环内数据类别一致性;若圆环内有数据与半径为 r 的圆内数据类别不一致,可缩小圆环宽度,再次甄别,直至数据符合同一类别要求。(5)分门别类。将确定好的数据分门别类进行标注,这样每位专家都可形成一个包含五种类别的数据集。(6)确定最终标记数据。为确保每位专家标记的数据符合统一的标准,可将各专家筛选形成的数据集,在符合同一类别的数据集基础上求取交集,这样就可去掉专家不同意见的数据,保留相同划分标准的数据,作为最终训练和检验深度学习模型的数据集。

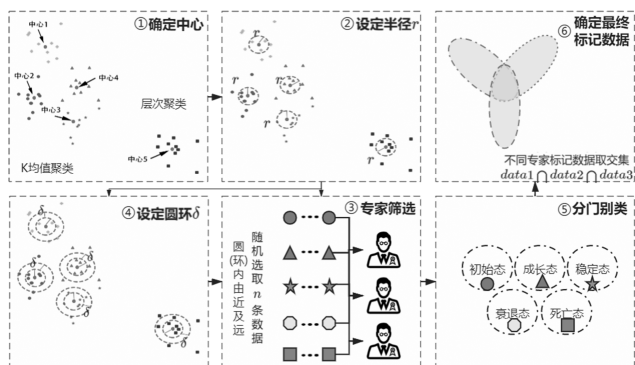


图3 网络学习资源进化态数据标注流程

(四)数学定义及模型设计

网络学习资源预警任务的目标是预测学习资源可能所处的进化状态,当学习资源有可能处于衰退态或死亡态时,及时给予资源开发方或维护方以预警信息。该任务数学定义如下:假设给定一个数据集 $D=\{(x_1, y_1), (x_2, y_2), \dots, (x_p, y_p)\}$,其中 $x_i \in R^n$ 表示学习资源的特征值,这些特征包括内容更新信息、资源关联信息、资源使用信息、资源批注信息、资源评分信息、资源受欢迎信息和资源采样时间间隔信息等内容。 $y_i \in \{0, 1, 2, 3, 4\}$ 表示学习资源的进化状态,0~5分别表示学习资源进化的初始态、成长态、稳定态、衰退态、死亡态。通过对网络学习资源的特征向量进行数学建模,寻找一种合适的函数 $y=f(x)$,实现对学习资源的状态进行预测。

在大规模实践中,可以利用深度学习模型自动提取数据特征的特性,完成对学习资源进化态预

测,以实现学习资源预警的目的。为此,本研究设计了一种基于卷积神经网络和 Word2Vector 结合的深度学习算法模型。通过卷积操作提取学习资源进化态之间的隐含信息,并尝试使用 Word2Vector 方法压缩 Id 的维度信息和提取资源进化过程的关联信息作为特征表示。具体的模型设计如图4所示,技术实现流程如下:

(1)将学习资源的 Id,通过 embedding 的方式进行编码,这是为了降低传统 one-hot 编码带来的维度灾难。本研究采用 Word2Vector 方式,将 Id 编码转换为稠密向量,再将该向量经过神经网络的方法转换为固定长度的特征向量,作为学习元 Id 的特征表示。上述过程数学表达式为(3-1)至(3-3)。

$$x'_i = \text{embedding}(x_i) \quad (3-1)$$

$$h_i = \text{Word2Vector}(x'_i) \quad (3-2)$$

$$u_i = \text{sigmod}(w_2 \cdot \text{sigmod}(w_1 \cdot h_i + b_1) + b_2) \quad (3-3)$$

其中, x'_i 表示学习元 Id 的随机初始化 embedding 之后的编码,Word2Vector()表示采用 Word2Vector 的方法, h_i 表示 Word2Vector 后获得向量, w_1, w_2, b_1, b_2 表示神经网络超参数,sigmod()表示 sigmoid 激活函数, u_i 表示最终提取的学习元特征表示向量。

(2)将 p 条学习资源的进化要素 n 维特征作为整体输入,通过卷积神经网络操作,提取学习资源进化态特征之间的潜在关系信息,再将提取后的特征信息通过神经网络转换为与 u_i 相同维度的特征向量 v_i ,学习资源进化态预测由 u_i 和 v_i 内积操作计算所得。上述过程数学表达式为(3-4)至(3-8)。

$$x'_i = \begin{bmatrix} W_{11} & \dots & W_{1n} \\ \dots & \dots & \dots \\ W_{p1} & \dots & W_{pn} \end{bmatrix} \quad (3-4)$$

$$c_i^j = f(W_c^j * x'_i(:, i:(i+ws-1))) + b_c^j \quad (3-5)$$

$$c^i = [c_i^1, c_i^2, \dots, c_i^j, \dots, c_i^{m-ws+1}] \quad (3-6)$$

$$d = [\max(c^1), \max(c^2), \dots, \max(c^j), \dots, \max(c^{nc})] \quad (3-7)$$

$$v_j = \tanh(W_2 \cdot \tanh(W_1 \cdot d + B_1) + B_2) \quad (3-8)$$

其中, x'_i 表示为输入的特征空间向量, c_i^j 表示经过第 j 个卷积核 W_c^j 操作之后提取的特征, $ws \in R^{m \times ws}$ 表示卷积窗大小, c^j 表示经过多个卷积核操作之后所得特征表示, $\max(c^j)$ 表示最大池化操作, d 表示经过多个最大池化操作之后提取的特征, W_1, W_2, B_1, B_2 表示神经网络超参数,tanh()表示双曲正切函数, v_j 表示最终提取的学习元进化要素的特征表示向量。

(3)进化态预测表达式为: $\hat{y}_i = u_i \cdot v_j$,通过进化态预测值 $\hat{y}_i$ 与真实值 y_i 之间误差函数计算误差,不断训练

深度学习模型,以获得最佳预测效果。

(4)误差函数选取多分类交叉熵损失函数计算误差,具体表达式见(3-9)。

$$E = \sum_{i=0}^4 \sum y_i \log(\hat{y}_i) \quad (3-9)$$

(5)模型训练方式采用 Adam 方式进行反向传播求解模型的超参数。

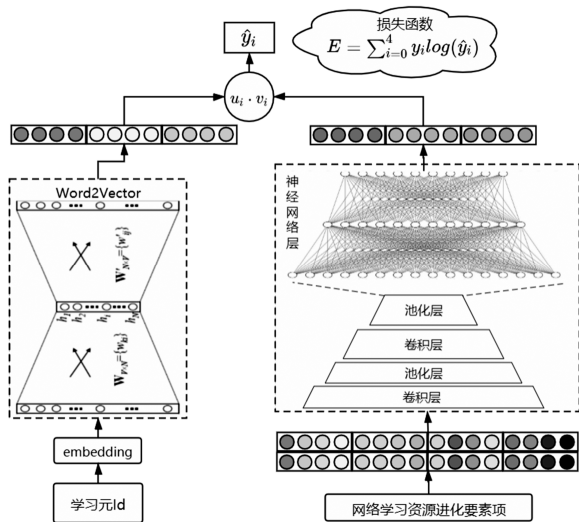


图4 网络学习资源进化预警模型

四、网络学习资源进化预警模型实验

为了客观地比较不同算法模型之间的性能,本研究主要从预测精确性角度比较不同模型的预测性能。研究挑选经典机器学习算法模型,如多层感知机(Multi-layer Perceptron,简称MLP)模型、支持向量机(Support Vector Machines,简称SVM)模型、神经网络模型(Neural Network,简称NN)、卷积神经网络(Convolutional Neural Networks,简称CNN)作为对比算法,与本研究设计的算法模型CNN+Word2Vec进行实验比对。

实验数据集为学习元平台采集的 learningcell-1k 和 learningcell-1M 两个数据集,其中 learningcell-1k 包含 3979 条数据,该数据集属于各进化态类别,数据分布较为平衡。learningcell-1M 包含 179392 条数据,该数据集属于各进化态类别,数据分布较不平衡,其中稳定态和衰退态数据占比较低。两个数据集的特征维度均包括学习元的进化要素项:学习元 Id、编辑数、版本数、浏览数、收藏数、批注数、打分数、准确性评分、客观性评分、完整性评分、规范性评分、及时性评分、点赞数、关联学习元数、关联知识群数、时间间隔(以周为单位)。且严格依据网络学习资源进化态标注

流程与方法,对两个数据集中的每条数据进化态进行了标记,标记值 $y \in \{0, 1, 2, 3, 4\}$ 。

实验采用 5 折交叉验证法,其中 80%数据用于模型训练,剩余 20%用于模型测试,选取 10 次实验的平均值作为最终的实验结果,以保证实验的公平性。实验评价主要选取精准率、F1 得分和 AUC (Area Under Curve)等三个评价指标,其中精确率主要反映预测为正样本中预测准确性占比,F1 分数则反映精确率和召回率的表现情况,AUC 为 ROC (Receiver Operating Characteristic Curve)曲线下的面积,用于反映分类效果好坏。

表2 各算法在 learningcell-1k 数据集的不同评价指标性能表现

评价指标	精确率	F1	AUC
对比算法			
SVM	0.6705	0.6691	0.6037
MLP	0.6625	0.6389	0.6173
NN	0.9127	0.9108	0.5956
CNN	0.9243	0.9326	0.6378
CNN+Word2Vec	0.9375	0.9592	0.6502

表3 各算法在 learningcell-1M 数据集的不同评价指标性能表现

评价指标	精确率	F1	AUC
对比算法			
SVM	0.6858	0.6906	0.7472
MLP	0.6307	0.5652	0.7564
NN	0.7699	0.7805	0.7769
CNN	0.8442	0.8252	0.7747
CNN+Word2Vec	0.8729	0.8641	0.7926

表2和表3表明本研究设计的算法在预测精准率、F1和AUC指标方面表现优于对比算法,说明了CNN+Word2Vec的算法模型框架可有效提升网络学习资源进化态预测的精准性。对比CNN与CNN+Word2Vec算法性能表现可以发现,双方在两个数据集的评价指标表现相近,但CNN+Word2Vec的算法设计可以提取数据间的隐含信息,故该算法表现效果略优于CNN算法。从数据分布来看,learningcell-1M数据集存在分布不平衡现象,因此需要更多地结合精准率、F1和AUC三个方面比较预测的精确性。

从精准率方面分析发现,CNN+Word2Vec算法对比其他算法有较大优势,与神经网络算法相比,在learningcell-1k数据集,本研究设计的算法精确率提升2.48%,而在learningcell-1M数据集,精确率提升10.3%。首先,这是因为卷积神经网络可以有效地提取

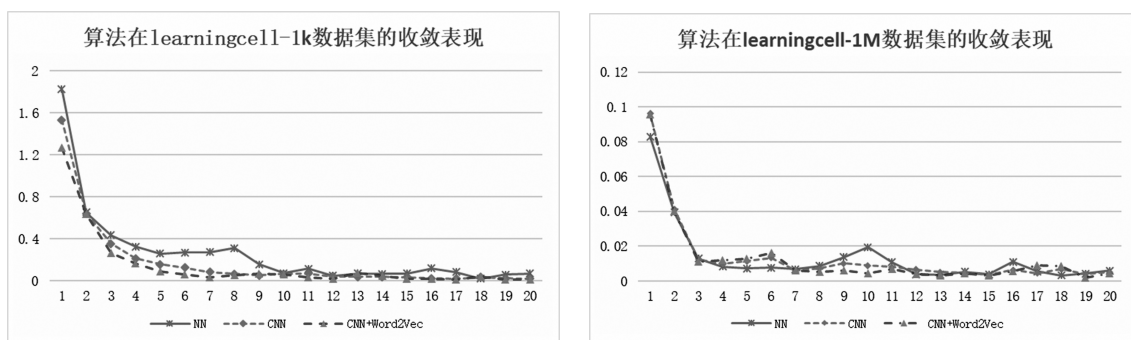


图5 算法在 learningcell-1k 和 learningcell-1M 数据集的收敛表现

输入特征之间的潜在信息,相对于神经网络而言,能更高效地数据化表示非线性关系。其次,不同数据及提升效果差异较大的原因是 learningcell-1M 数据分布不平衡现象相较于 learningcell-1k 严重,这也反映了本研究设计模型的抗干扰能力较强,在数据分布不均衡时,仍可以有效进行网络学习资源进化态的预测。

从 F1 得分方面分析发现,本研究设计的算法模型表现优于对比算法,从综合精确率和召回率两方面看,CNN+Word2Vec 算法在预测学习资源进化态时更为有效,并且当数据分布越均衡时,算法模型的表现效果也越好。

从 AUC 方面分析发现,CNN+Word2Vec 算法对比 CNN 算法提升 2%,这表明本研究设计的算法模型在预测全面性方面优于对比算法,纵向对比 learningcell-1k 和 learningcell-1M 算法的 AUC 可以发现,learningcell-1M 的 AUC 表现较好,这是因为 learningcell-1M 的数据量较多于 learningcell-1k 的数据量,大部分数据集中在初始态类别中,模型能学习较多的数据信息,从而提升预测分类的效果。

从图 5 分析可以发现,CNN+Word2Vec 算法在数据集 learningcell-1k 和 learningcell-1M 收敛效果表现比 CNN 和 NN 算法的收敛效果更好。CNN+Word2Vec 算法在数据集 learningcell-1k 和 learningcell-1M 的收敛速度比神经网络算法收敛更快,这样有利于预警算法模型在数据集中快速找到最优参数,在实际应用中也可快速地进行模型的优化。CNN+Word2Vec 算法在两个数据集的收敛平稳性表现都优于 CNN 和 NN 算

法,这也表明算法在求解过程中较为稳定,在实际应用过程有利于网络学习资源进化预警的稳定性。

五、结 语

本研究以网络学习资源进化预警模型构建为核心,探讨了资源进化态表征、资源进化要素项提取、资源进化态标注、资源进化预警任务的数学定义等问题,并以学习元平台采集数据为实验数据集,设计并实现了卷积神经网络+Word2Vec 的网络学习资源进化预警算法模型。从与经典算法对比实验效果来看,本研究设计的算法模型相较于经典的机器学习算法和神经网络算法,在预测精准率、F1 和 AUC 等指标方面表现具有一定的优势,且模型在训练过程中收敛态势也优于对比算法,相信在未来的实际应用中能发挥更佳的预警效果。

网络学习资源进化预警技术模型的设计与实现,既是对网络学习资源进化预警技术的新探索,也可帮助各地智慧教育云平台提升海量网络学习资源的管理效能。目前本研究还缺乏在实际平台的应用效果反馈,在预警算法层面还有诸多可改进之处。例如,可以尝试在预警模型中引入时间序列信息和评论数据信息,以实现更加精准、科学的网络学习资源进化预警。也可尝试借鉴前沿的可解释性深度学习技术,研发具有可解释性的网络学习资源进化预警信息生成技术模型,实现在资源进化预警的同时,智能生成具体的预警内容信息,辅助资源维护人员实现网络学习资源的及时精准优化。

[参考文献]

- [1] 杨现民,余胜泉.泛在学习环境下的学习资源进化模型构建[J].中国电化教育,2011(9):80-86.
- [2] YU S Q, YANG X M, CHENG G, etc. From learning object to learning cell: a resource organization model for ubiquitous learning[J]. Journal of educational technology & society, 2015, 18(2): 206-224.
- [3] YANG X M, QIU Q, YU S Q, etc. Designing a trust evaluation model for open-knowledge communities [J]. British journal of educational technology, 2014, 45(5): 880-901.

- [4] ATTIYA H, BURCKHARDT S, GOTSMAN A, etc. Specification and complexity of collaborative text editing [C]. In Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing. Chicago, IL, USA, 2016: 259–268.
- [5] SCERBAKOV N, KAPPE F, PAK V. Collaborative document authoring as an e-learning component [C]. In EdMedia+ Innovate Learning, Association for the Advancement of Computing in Education (AACE). Las Vegas, NV, USA, 2018: 101–107.
- [6] 余胜泉,陈敏.泛在学习资源建设的特征与趋势——以学习元资源模型为例[J].现代远程教育研究,2011(6):14–22.
- [7] 杨现民.泛在学习环境下的学习资源有序进化研究[J].电化教育研究,2015,36(1):62–68.
- [8] 徐刘杰,余胜泉.泛在学习资源进化的动力模型构建[J].电化教育研究,2018,39(4):52–58.
- [9] 杨现民,余胜泉.开放环境下学习资源内容进化的智能控制研究[J].电化教育研究,2013,34(9):83–88.
- [10] 何向阳.用户参与网络信息资源再生影响因素的实证研究[J].远程教育杂志,2015(5):90–98.
- [11] LIU S, BREMER P T, THIAGARAJAN J J, etc. Visual exploration of semantic relationships in neural word embeddings[J]. IEEE transactions on visualization and computer graphics, 2018, 24(1):553–562.
- [12] HERMINIO G G, JOSÉ E L, PAULE R M. Enhancing e-Learning content by using semantic web technologies[J]. IEEE transactions on learning technologies, 2017, 10(4):544–550.
- [13] MUSTO C, NARDUCCI F, LOPS P, etc. Linked open data –based explanations for transparent recommender systems [J]. International journal of human-computer studies, 2019(121):93–107.
- [14] 杨现民,余胜泉,张芳.学习资源动态语义关联的设计与实现[J].中国电化教育,2013(1):70–75.
- [15] GASPARETTI F, DE MEDIO C, LIMONGELLI C, etc. Prerequisites between learning objects: automatic extraction based on a machine learning approach[J]. Telematics and informatics, 2018, 35(3):595–610.
- [16] SHEN Y, DING N, ZHENG H, LI Y, YANG M. Modeling relation paths for knowledge graph completion [J]. IEEE transactions on knowledge and data engineering, 2021,33(11):3607–3617.
- [17] 王怀波,陈丽.网络化知识的内涵解析与表征模型构建[J].中国远程教育,2020(5):10–17,76.
- [18] MARBOUTI F, DIESFES-DUX H A, MADHAVAN K. Models for early prediction of at-risk students in a course using standards-based grading[J]. Computers & education, 2016(103):1–15.
- [19] DANIELE D M, MAREN S, HENDRIK D, etc. Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data [C]. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference, ACM, Vancouver British Columbia Canada, 2017: 188–197.
- [20] MARQUEZ-VERA C, CANO A, ROMERO C, etc. Early dropout prediction using data mining: a case study with high school students[J]. Expert systems, 2016,33(1):107–124.
- [21] KOTSIANTIS S B. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades[J]. Artificial intelligence review, 2012, 37(4):331–344.
- [22] HU Y H, LO C L, SHIH S P. Developing early warning systems to predict students' online learning performance [J]. Computers in human behavior, 2014(36):469–478.
- [23] 李景奇,卞艺杰,方征.基于 BKT 模型的网络教学跟踪评价研究[J].现代远程教育研究,2018(5):106–114.
- [24] YANG T Y, BRINTON C G, JOE-WONG C, CHIANG M. Behavior-based grade prediction for MOOCs via time series neural networks[J]. IEEE journal of selected topics in signal processing, 2017, 11(5):716–728.
- [25] 陈晋音,方航,林翔,等.基于在线学习行为分析的个性化学习推荐[J].计算机科学,2018(S2):422–426.
- [26] FEI H, LI H. The study of learners' emotional analysis based on MOOC [C]. In Proceedings of the 2018 IEEE International Conference on Cognitive Computing, San Francisco, USA, 2018: 170–178.
- [27] WANG L, HU G, ZHOU T. Semantic analysis of learners' emotional tendencies on online MOOC education[J]. Sustainability, 2018, 10(6):1921–1940.
- [28] WANG X, ZHAO M, HUANG C, etc. Predicting students' mood level using multi-feature fusion joint sentiment-topic model in mobile Learning [C]. In the 21st ACM Conference on Computer Supported Cooperative Work and Social Computing, New York, USA, 2018: 316–330.

Research on Early Warning Model for Evolution of E-learning Learning Resource

YANG Xianmin, YUAN Meng, LI Kangkang

(Jiangsu Engineering Research Center of Educational Informationization, Jiangsu Normal University,
Xuzhou Jiangsu 221116)

[Abstract] Large-scale open collaborative environment has brought new challenges to the efficient management of e-learning resources. The traditional management methods based on page views and rating rankings are difficult to guarantee the generation and continuous evolution for high-quality e-learning resources. How to construct an intelligent early warning model for the evolution of e-learning resources is the core problem of solving the efficient management of e-learning resources in the era of big data. Based on the implementation process of e-learning resource evolution warning task, this study constructs a technical model of e-learning resource evolution warning on the basis of exploring the issues of resource evolution state representation, resource evolution element item extraction, resource evolution state labeling, and mathematical definition of resource evolution warning task. The experimental performance on two datasets extracted from the Learning Cell Platform shows that the proposed early warning model outperforms the comparison algorithm in terms of predicting accuracy, F1, AUC and other indicators.

[Keywords] E-learning Resource; Learning Resource Evolution; Early Warning Model; Deep Learning; Convolutional Neural Network

(上接第 46 页)

to emphasize the learning process, and the open connection organization system based on learners' needs. In addition, the multi-dimensional relationships among the system components are also sorted out from formal logic, structural logic, hierarchical logic and functional logic. The construction of the large-scale personalized teaching system requires multi-agent participation to cultivate a collective intention. It is necessary to train teachers' ability to implement the institutional function step by step, to form a collaborative rule system through the construction of auxiliary institutions, and to improve the indirect system and nurture the overall ecology.

[Keywords] Large-scale Personalized Teaching; Big Data; Online Education; Teaching Institution; Personalization