文章编号:1674 - 2869(2018)01 - 0109 - 05

MOOC学习中“伪学习者”行为特征分析与识别的研究

王传安[1,2],葛 华[1]

1. 安徽科技学院信息与网络工程学院,安徽 滁州 233100;

2. 北京邮电大学网络技术研究院,北京 100876

摘 要:为深入了解学习者在慕课(Massive Open Online Courses,MOOC)平台下的学习行为特性,识别可能存在的伪学习行为,提出了基于协同训练的伪学习者识别模型。首先在学习者概要特征基础上,提出自主行为特征和交互信息特征,并将三者联合优化分析,以建立学习者动态行为模式;然后采用多分类器协同学习的方法对学习者行为数据进行标记,根据标记结果以判定学习者是否为伪学习者。最后的数据表明,基于协同训练的伪学习者识别模型能有效地判别一个学习者是否是伪学习者,为今后检测MOOC教学效果提供了一种依据。

关键词:慕课;伪学习者检测;学习行为;协同训练

中图分类号:G434 **文献标识码**:A **doi**:10. 3969/j. issn. 1674-2869. 2018. 01. 020

Analysis and Detection of Pseudo-Learners Behavior in Massive Open Online Courses

WANG Chuan'an[1,2], *GE Hua*[1]

1. Anhui Science and Technology University, Chuzhou 233100, China;

2. Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: To detect the pseudo-learners in Massive Open Online Courses (MOOC), a pseudo-learner detection model was proposed based on Tri-training. First, new behavior characteristics such as autonomic behavior and interactive information were proposed based on the analysis of the learners´ fundamental characteristics, and the learner's dynamic behavior pattern was established by the joint analysis of the above characteristics. Second, a multi-classifier was used to identify the learner's behavior data to detect pseudo-learners. Finally, experiment results indicate the proposed method can be used to detect whether a learner is a pseudo-learner, which has the potential to partially evaluate the MOOC teaching effect in practice.

Keywords: MOOC; pseudo-learners detection; learning behavior; Tri-training

慕课(massive open online courses,MOOC)作为一种新兴的学习者协同进行的学习平台,随着互联网Web 2.0和云计算等技术的成熟[1],不但吸引了越来越多学习者、平台提供者及高校参与其中,同时也引发了教育研究者的极大关注[2]。尽管MOOC得到了快速的发展,一些初步的研究成果已经形成,但是教师对学习者在MOOC平台下是如何学习的知之甚少[3-4]。了解学习者的学习行为特性,不仅可以优化MOOC平台的设计,更重要的是可以弥补MOOC教学方式中师生的时空分离缺点。

现在对MOOC学习者学习行为的研究多数倾向于学习行为方式与学习效果之间的关联[5-6],且研究所有的样本数据多数直接采集行为日志或问卷调查,属于线后统计分析行为[7]。同时,目前对

收稿日期:2017-04-04

基金项目:安徽省质量工程教研项目(2014Mvoco38)

作者简介:王传安,博士研究生,副教授。 E-mail:wangchuanan1010@126.com

引文格式:王传安,葛华. MOOC学习中“伪学习者”行为特征分析与识别的研究[J]. 武汉工程大学学报,2018,40(1):109-113.

MOOC学习者行为分析都是单方面的，无法判定学习者在学习过程中是否存在伪学习行为。比如某学习者为尽快播放完课程教学视频，采取连续播放或挂机方式播放教学视频，且在播放过程中多次拖放视频。

再比如某学习者为完成提交课程作业任务，一次性提交所有作业，诸如此类学习行为我们称之为伪学习行为，同时该学习者可看做是“伪学习者”[8-9]。而如何正确识别MOOC平台中的伪学习者，并根据伪学习行为特性制定相应的预防措施或报警机制，以抑制学习者的伪学习行为，已成为MOOC教学研究中首要解决的关键问题之一。

本文为实现高精度的伪学习者识别，将多种学习行为特征进行融合分析，建立了学习者动态行为模式，通过多分类器协同学习，实现对未标记学习者行为数据进行标记，进而根据标记判定该学习者是否为伪学习者。

1 学习者动态行为模式

由于仅基于某一特定类型的行为特征并不能够对伪学习者进行准确的识别，因此，在原数据集中提供的学习者概要特征基础上，提出学习者自主行为特征和学习者交互信息特征，并将其三者融合以对学习者动态行为模式进行建模。

1.1 学习者概要特征

学习者自身的因素，如性别、文化程度、选课时间和学习目的，是影响学习者学习行为发生的内部因素，文中将这些因素作为伪学习者识别的概要特征。

1.2 学习者自主行为特征

1.2.1 *课程关注度* 在实际的MOOC学习过程中，伪学习者自从注册后很少登录MOOC平台，即使经常登录平台，却采用走马观花的学习方式对待学习内容，导致每次学习时长较短，故与正常学习者相比，伪学习者的平台登录次数和学习时长都低的多。因此，为了衡量学习者自愿学习某课程的意愿程度，定义了用户课程关注度 F_u 为

$$F_u=\frac{\sum_{j=1}^{N}t_{u,j}}{D*T} \tag{1}$$

式(1)中 $t_{u,j}$ 是学习者 u 第 j 次登录MOOC平台时的学习时长；N 为学习者 u 登录的总次数；而 D 和 T 分别表示学习该课程时的建议学习天数和每次学习的时长。

1.2.2 *视频学习行为* 教学视频是MOOC平台中最主要的学习资源，同时也是学习者获取知识的最主要途径[1]。伪学习者为急于结束视频学习，常采用“挂机”的方式，在一次登录过程中持续播放整个课程的所有教学视频。因此，本文将观看教学视频的频度熵作为识别伪学习者“挂机”视频学习行为的特征量。学习者 u 的视频观看频度熵 E_u 定义为：

$$E_u=\frac{1}{\ln D}\sum_{d=1}^{D}\frac{l_{u,d}}{K}\ln\frac{l_{u,d}}{K} \tag{2}$$

式(2)中 $l_{u,d}$ 是学习者 u 在第 d 天学习的教学视频数量，K 是学习者 u 需要学习的某门课程总的教学视频个数。对于一个学习者，如果他将要学习的课程教学视频在某一天观看完，则他的视频观看频度熵趋于0；如果他将所有课程教学视频平均分布在 D 天观看完，则视频观看频度熵为1。因此，较高的频度熵值代表了学习者有规律地观看课程视频，能较好的按照视频学习建议进行学习。

另一个与视频学习行为相关的特征量是观看视频时的行为动作。在观看教学视频时，正常学习者的主要动作包括暂停、后退及个别快进动作，而伪学习者带着一颗“应付的心”的观看教学视频，其动作主要是快进及拖拽。采用方差来衡量正常学习者与伪学习者在观看视频时的行为动作差异：

$$P_u=\frac{\sum_{k=1}^{K}((q_{u,k}+s_{u,k})-(\bar{q}_k+\bar{s}_k))^2}{K} \tag{3}$$

式(3)中 $q_{u,k}$ 表示学习者 u 在第 k 个教学视频上快进的次数，$s_{u,k}$ 表示学习者 u 在第 k 个教学视频上拖拽的次数，$\bar{q}_k$ 和 $\bar{s}_k$ 分别表示所有学习者在第 k 个教学视频上快进的次数和拖拽次数的平均值。对于一个学习者 u，其动作差异 P_u 越大，说明该学习者的快进和拖拽次数越多，反之亦然。

1.3 学习者交互信息特征

1.3.1 *动态发帖数* 有时为了制造主动交互学习的假象，伪学习者在一次登录过程中在讨论区发布多个帖子，或者在每次登录过程中都发布多个讨论帖[10]。因此，为了衡量发帖数与登录次数间的关系，将发帖数与登录发帖区次数的比率作为动态发帖数的特征表示：

$$\text{ratio\_}R_u=\frac{\sum_{j=1}^{N}r_{u,j}}{N} \tag{4}$$

式(4)中 $r_{u,j}$ 是学习者 u 第 j 次登录MOOC平台时在讨论区发帖的个数。

1.3.2 *发帖内容相关性* 在MOOC学习中，一个正常学习者针对不同的教学内容，发布的多个帖子在内容和主题上并不会表现出很强的自相似

性，而伪学习者一般采用内容模板在一次登录过程中发布大量具有较高相似度的帖子。因此，本文从文本角度出发，衡量用户 u 发帖内容的相关性，其计算公式如下：

$$\mathrm{sim}_u = \sum_{w=1}^{W} \frac{\Gamma(w, w-1)}{T(w, w-1)} \tag{5}$$

式(5)中 W 表示用户 u 发布的帖子总数，$T(w, w-1)$ 表示第 w 条帖子与第 $w-1$ 条帖子间的发布时间间隔，$\Gamma(w, w-1)$ 表示第 w 条帖子与第 $w-1$ 条帖子间的jaccard相似度[11]，其计算公式为：

$$\Gamma(w, w-1) = \left| \frac{G(w) \cap G(w-1)}{G(w) \cup G(w-1)} \right| \tag{6}$$

其中，$G(w)$ 和 $G(w-1)$ 分别表示第 w 条帖子与第 $w-1$ 条帖子中所包含的相似词语集合。

1.3.3　*动态作业数*　伪学习者为了完成MOOC学习过程中的作业提交任务，往往在一次登录过程中将多个教学内容环节的不同作业提交到系统平台中。与动态发帖数类似，采用提交作业数与登录次数的比率作为动态作业数的特征表示为：

$$\mathrm{ratio}\_H_u = \frac{\sum_{j=1}^{N} h_{u,j}}{N} \tag{7}$$

式(7)中 $r_{u,j}$ 是学习者 u 第 j 次登录MOOC平台时提交的作业个数。

2　伪学习者识别

由于仅基于某一特定类型的特征并不能够对伪学习者进行准确的识别，因此，融合用户概要特征、用户关系特征以及用户发布信息特征，通过多分类器协同学习，实现对未标记学习者行为数据进行标记，进而根据标记判定该学习者是否为伪学习者。其识别模型如图1所示。

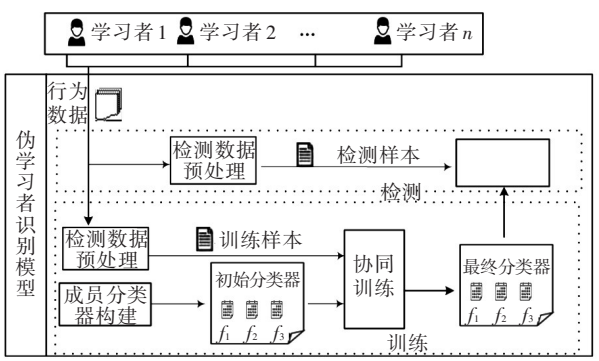

图1　基于协同训练的伪学习者识别模型

Fig. 1　Pseudo-learner recognition model based on collaborative training

2.1　数据预处理

采集MOOC平台中的学习者学习行为数据，根据伪学习者行为特征指标，提取每条行为数据的特征值，构造供集成分类器使用的训练样本集和检测样本集。由于协同学习采用SVM进行训练分类，而SVM只能处理数值型的数据，为此要对所提取特征值进行标准化和离散化处理。

为提高分类器性能，协同学习在训练过程中需要利用已标记样本和未标记样本对多个分类器进行协同训练。因此，为获得已标记样本，对预处理后的学习者行为数据进行类别标识，用lab表示，标记lab取值为{1，-1}，当值为1表示该样本为正常学习者行为，而-1表示该样本为伪学习者行为。具有 n 个特征值的训练样本可表示为 $X=[(x_1, x_2, \cdots\cdots x_n), \mathrm{lab}]$，其中 x_i 为学习者第 i 个学习行为特征的取值，其中 $i \in [1, n]$。

2.2　伪学习者识别

文中应用选择性协同学习解决伪学习者的识别检测问题[12]，其识别过程可划分为协同训练阶段和检测识别阶段。

2.2.1　*协同训练*　根据学习者不同的学习行为特征，将已标记的学习者行为样本分为3个无重复的训练集，然后使用3个训练集分别训练初始分类器 f_1、f_2 和 f_3，3个分类器协同工作，遇到未标记行为样本时，若3个分类器对该样本预测的标记一致时，使用预测标记对该样本进行标记，并将已标记的新样本添加到已标记样本集中，形成新的训练集，开启新一轮迭代训练，直至分类器不再发生变化。

2.2.2　*伪学习者检测*　分类器训练结束后，在未标记的学习者行为样本监测与分类识别中，分类器采用协同投票方法对学习者行为样本的类别进行判定，若判定样本标记lab值为-1，则认定该样本为伪学习者。根据陈文等[13]提出的样本检测错误率判定方法及Angluin等[14]提出的噪声学习理论，样本检测错误率 $\Im$ 与分类数目 B 间的关系应满足式(8)：

$$(1-2\Im)^2 > \frac{2}{\ell^2 A} \ln\left(\frac{2B}{\sigma}\right) \tag{8}$$

其中 A 为样本个数，σ 为置信度参数，ℓ 是检测错误率上限。将式(7)进一步转换得式(9)：

$$\Im \leqslant \frac{1 - \frac{2}{\ell^2 A} \ln\left(\frac{2B}{\sigma}\right)}{4} \tag{9}$$

令 $\hbar = 2\tau \ln\left(\frac{2B}{\sigma}\right)$，其中 τ 为使式(8)取等号的系数，可得式(10)：

$$A(1-4\Im) = \frac{\hbar}{\ell^2} \tag{10}$$

设检测第 u 个未标记的样本数据时，样本检测

错误率为$\Im^{u-1}$，若满足式(11)：

$$(A+1)(1-4\Im^{u})=A(1-4\Im^{u-1}) \tag{11}$$

则表明增加第x个样本后能改进分类器性能，这也意味着对第u个样本预测的标记是精确的；否则放弃本轮对检测样本x的类别判定，从检测样本集中重新进行选择样本，进行下一轮的检测。式(11)中$A+1$表示将第x个样本加入已标记样本集后的规模，$\Im^{u}$为完成第x个样本检测后的检测错误率。

3 实验分析

实验所使用的学习者学习行为日志数据均来自于MOOC课程《大学计算机基础》，对采集到的行为数据进行特征抽取，并按学号(SID)进行分类排序，然后根据文中第二部分动态行为模式建模所需要的行为特征进行格式处理和离散化处理，表1为处理后的部分数据样本实例(表1中F_u为课程关注度、E_u为视频观看频度熵、P_u为观看视频行为动作特征、R_u为动态发帖特征、Sim_u为发帖内容相关性、H_u为动态作业特征)。

表1 处理后的学习行为数据样本

Tab. 1 Samples of processed learning behavior data

SID	F_u	E_u	P_u	R_u	Sim_u	H_u
S1600101	0.95	0.92	0.03	1.5	0.25	1
S1600102	0.85	0.62	0.25	0.7	0.03	1.05
S1600105	0.75	0.85	0.65	0.89	0.05	0.95
S1600107	0.92	0.52	0.54	0.95	0.35	0.77
S1600108	0.45	0.55	0.35	0.8	0.45	1.65
S1600201	0.15	0.05	0.9	4.5	0.75	5.8
S1608501	0.55	0.75	0.68	0.73	0.48	0.89
…	…	…	…	…	…	…
S1609001	0.8	0.02	0.03	6.5	0.89	0.25

在伪学习者预测效果的评价方面，本文采用准确率和召回率作为评价指标。准确率描述了分类器将正常学习者与伪学习者正确分类的百分比，而召回率表明了检测出的伪学习者中，真实伪学习者的比率[15]。表2记录了两组实验样本集的实验结果，其中样本集1中共5 000条样本数据，其中3 000条作为训练数据，2 000条作为测试数据；而样本集2中共3 000条样本数据，其中1 500条作为训练数据，另外1 500条作为测试数据。

在采用样本集1进行实验时，采用本文提出的6个行为特征训练分类器，而采用样本集2进行实验时，添加了年龄、选课时间和性别三个特征训练分类器。根据表2的实验结果，发现并不是行为特征选取的越多，预测效果越好，因为部分特征具有“负效果”，反而会降低分类器的准确率。这也证明了所提出的动态行为模式的有效性。

表2 伪学习者预测结果

Tab. 2 Predict results of pseudo-learners %

数据集	准确率	召回率
样本集1	97.85	98.5
样本集2	97.78	98.45

图2给出了所有学习者的视频观看频度熵曲线。从图2中可以看出极少数学习者的视频观看频度熵趋于0，表明这些学习者的视频观看行为特别集中，极有可能是采用挂机播放的方式观看教学视频，而大部分学习者的视频观看频度熵都在0.5以上，表明他们的视频观看行为分布较为平均。

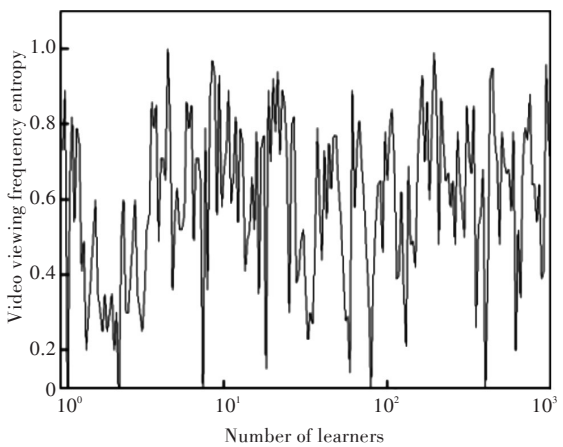

图2 学习者观看视频行为统计

Fig. 2 Video viewing behavior statistics of learners

表3进一步给出了正常学习者和伪学习者的视频观看频度熵的对比，只有不超过10%的正常学习者的频度熵小于0.25，而伪学习者的比例高达95%。这说明伪学习者为了应付学习，在一次或几次登录过程中采用连续播放(或挂机播放)的方式将课程教学视频播放完，即与正常学习者相比，伪学习者的视频观看行为特别集中。

表3 正常学习者与伪学习者的视频观看频度熵

Tab. 3 Video viewing frequency entropy of normal learners VS pseudo-learners

频度熵	$>E_u$的比例	
	正常学习者	伪学习者
0.05	0	0.65
0.1	0.02	0.7
0.15	0.05	0.85
0.2	0.07	0.9
0.25	0.1	0.95

从图3的统计结果可以看出伪学习者的发帖内容的相似度高于正常学习者。在MOOC学习中,一个正常学习者针对不同的教学内容,在讨论区发帖的内容一般会与教学内容紧密相关,因此发帖内容相关性程度较低,而伪学习者一般采用内容模板在一次登录过程中发布大量具有较高相似度的帖子。

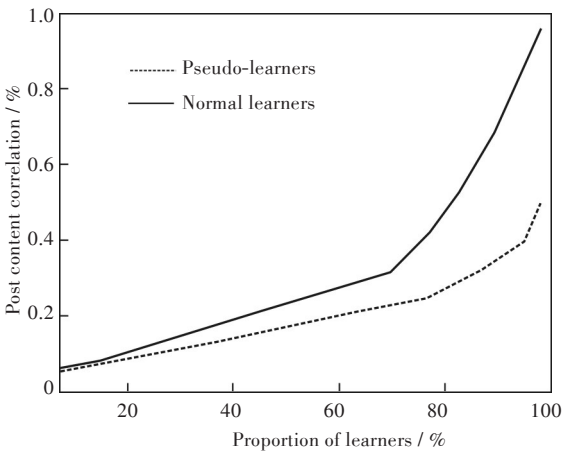

图3　伪学习者与正常学习者发帖相关性比较

Fig. 3　Post content correlation of normal learners VS pseudo-learners

4　结　语

以MOOC环境下学习者的学习行为作为研究落脚点,根据学习者学习行为特性,对学习者动态行为模式进行建模,然后采用协调训练的方法进行学习行为数据训练,以此来检测学习过程中是否存在伪学习行为。为验证文中所提的伪学习者识别方法的有效性,选取了MOOC平台中《大学计算机基础》作为研究实例,将采集到学习者学习行为日志数据作为实验样本,采用分类标记的方法进行伪学习者识别验证。结果证明,文中所提的伪学习者识别方法具有较高的准确率和召回率。

致谢

在此对文中实验和测试等提供支持和帮助的安徽科技学院《大学计算机MOOC平台》研究组各位老师表示感谢。

参考文献:

[1]　蒋卓轩,张岩,李晓明. 基于MOOC数据的学习行为分析与预测[J]. 计算机研究与发展,2015,52(3):614-628.

[2]　梁林梅. MOOCs学习者分类特征与坚持性[J]. 比较教育研究, 2015, 37(1):28-34.

[3]　CHANG J W. Exploring engaging gamification mechanics in massive online open courses [J]. Journal of Educational Technology & Society, 2016 , 19 (2) : 177-203.

[4]　李帅,张岩峰,于戈,等. MOOC平台学习行为数据的采集与分析[J]. 中国科技论文, 2015, 10(20): 2373-2376.

[5]　RODRIGUEZ C. MOOCs and the AI-stanford like courses: two successful and distinct course formats for massive open online courses [J]. European Journal of Open, Distance and E-Learning , 2012, 1(2):1-13.

[6]　BRESLOW L, PRITCHARD D , DEBOER J, et al. Studying learning in the worldwide classroom research into edX's first MOOC [J]. Research & Practice in Assessment, 2013 , 8 (1):13-25

[7]　MILLIGAN C, LITTLEJOHN A , MARGARYAN A. Patterns of engagement in connectivist MOOCs [J]. Journal of Online Learning & Teaching, 2017, 9(2): 149-159.

[8]　SHEN C W, KUO C J. Learning in massive open online courses: Evidence from social media mining [J]. Computers in Human Behavior , 2015 , 51(3):568-577.

[9]　GLYN H, CHELSEA D. The utilization of data analysis techniques in predicting student performance in massive open online courses (moocs) [J]. Research and Practice in Technology Enhanced Learning, 2015, 10 (1):1-18.

[10]　HEATHER B, SHAPIROC C, NOELLE E, et al. Understanding the massive open online course (MOOC) student experience: an examination of attitudes, motivations, and barriers [J]. Computers & Education , 2017 , 110(3) :35-50.

[11]　WANG M M,ZUO W L,WANG Y. A multidimensional nonnegative matrix factorization model for retweeting behavior prediction [J]. Mathematical Problems in Engineering Volume, 2015,5(1): 1-10.

[12]　陆悠,李伟,罗军舟,等. 一种基于选择性协同学习的网络用户异常行为检测方法[J]. 计算机学报, 2014,37(37):28-40.

[13]　陈文,张恩阳,赵勇. 基于多分类器协同学习的卷积神经网络训练算法[J]. 计算机科学,2016,43(9): 223-227.

[14]　ANGLUIN D,LAIRD P. Learning from noisy examples [J]. Machine Learning, 1988, 2(4): 343-370.

[15]　李赫元,俞晓明,刘悦,等. 中文微博客的垃圾用户检测[J]. 2014,28(3):62-68.

本文编辑:陈小平