# 网络学习空间中学习画像的标签模型构建研究

赵　春,李　欣
(成都锦城学院 计算机与软件学院,四川 成都 611731)

**摘　要**:混合式学习是互联网时代的一种主要学习形式。学生在网络学习空间中的学习活动将会产生大量的学习过程和学习结果数据。挖掘这些数据的价值成为了教育信息化的热点问题。基于这些学习数据开展学生学习画像将有助于分析学生的学习行为和学习成效,为探索个性化教育提供充分的数据支撑。该文以学习画像中最典型的学习能力和学习行为偏好两个维度的分析刻画为例,根据学习数据的动态周期性特征,提出了一种调整的线性加权变异系数算法,消除了量纲与样本容量的影响,实现了学生学习能力稳定性标签模型;同时利用箱线图 $k$ 百分位数方法,结合偏好随机变量概率分布理论,构建了学生行为偏好中最有代表性的学习任务响应习惯标签模型。通过对采集的样本数据和处理结果的对比分析表明,采用该方法构建的两种标签模型具有良好的刻画效果,准确地反映了学生的个体学习特征。

**关键词**:学习画像;学习能力;学习行为;响应习惯;标签模型;变异系数;箱线图
**中图分类号**:TP399　　**文献标识码**:A　　**文章编号**:1673-629X(2023)10-0176-07
doi:10.3969/j.issn.1673-629X.2023.10.027

# Research on Building Label Model of Learning Profile in Network Learning Space

ZHAO Chun, LI Xin
(School of Computer and Software, Chengdu Jincheng College, Chengdu 611731, China)

**Abstract**: Blended learning is a major learning method in the age of network. Students' learning activities in network learning space will produce a large amount of learning process and learning result data. Mining the value of these data has become a hot issue of education informatization. Portraying student learning based on these learning data will be helpful to analyze students' learning behavior and learning effectiveness, and then provide sufficient data support for exploring personalized education. Learning ability and learning behavior preference are the most typical dimensions in the learning profile. Taking the two dimensions of learning ability and learning behavior preference as an example, according to the dynamic periodic characteristics of learning data, we propose an adjusted linear weighted coefficient of variation algorithm, which eliminates the influence of dimension and sample size, and realizes the stability label model of students' learning ability. Using boxplot, k-percentile and probability distribution theory of preference random variables, the most representative learning task response habit label model is constructed. By comparative analyzing of the collected sample data and processing results, the two label models have excellent characterization effects, and accurately reflect the individual learning characteristics.

**Key words**: learning profile; learning ability; learning behavior; response habits; label model; variation coefficient; boxplot

## 1　概　述

在“互联网+教育”的背景下,随着移动智能设备的普及和数字化学习资源的极大丰富,网络学习逐渐成为一种主流的学习模式。学生在网络学习空间中的学习行为产生了大量的学习数据。利用基于大数据的用户画像技术对学生的线上学习数据进行挖掘分析、构建学生学习画像变得现实可行。

用户画像是根据用户数据提炼出的描述用户属性及行为的标签集合[1],被广泛地应用于描述用户特征、用户兴趣和用户偏好等[2-4]。学生画像则是用户画像技术在教育领域的应用,反映了学生的学习特征和学习行为。它可以帮助教师理解教学实施情况,也可以辅助制定新的教学策略[5]。余明华等将学生画像划分为能力属性、行为属性和兴趣属性,以数据分析和人工手段相结合的方式建立了学生画像的标签体系[6]。杨长春等认为创建用户画像的过程就是依据构建的用户模型在用户信息中得到特征,并将特征标签化的过程[7]。他从学生的基本特征、学习特征、学习能力、素

**收稿日期**:2022-12-26　　**修回日期**:2023-04-27
**基金项目**:四川省高等教育学会 2021 年教育信息化研究课题(GJXHXXH21-YB-29)
**作者简介**:赵　春(1978-),男,硕士,副教授,通讯作者,研究方向为大数据技术、教育信息化。

质与偏好五个维度进行了学生画像建模。黄文林认为学生画像是用能够反映学生的特征描述、行为诊断和需求预测属性的三类标签来刻画,并进行可视化呈现的用户画像方法[8]。任红杰认为学生画像是根据学生的基础信息、学习习惯、学习偏好、学习行为和学习期待等方面的数据信息构建出来的标签化学生模型[9]。杨彩霖认为可以从线上学习的活跃度、参与度、持久度、学习效果和学习预警五个维度刻画学生个体画像,并对每个维度赋予相应的权值[10]。

以上研究基于各自不同的数据基础和画像需求,从不同的角度提出了构建学生画像标签模型的方法。它们各自抽取的数据维度和粒度虽然有所不同,但学习能力和学习行为均被包含其中,是最被研究者重视的两个维度。上述研究中提到的学习习惯和学习偏好等维度完全可以合入学习行为维度中体现。学习能力标签模型可以以学生的学习成绩为主要依据进行分析刻画,而学习行为标签模型的构建所依赖的数据维度则相对较为复杂,比如设备使用习惯、登录时间习惯、作业完成习惯和学习响应习惯等。

在构造学习画像标签的过程中,传统方式采用的单纯统计类标签维度刻画的模式具有颗粒度粗糙、标签等级不够精准的缺陷。因此很多研究者利用聚类方法进行用户分类与画像构建。张毅认为大数据背景下用户画像的统计方法可以简单概括为针对用户属性加以统计,建议从统计分析视角出发,明确画像指标,做好主客观指标之间的转换,从而获得用户画像更详细的特征[11]。翟鸣宇等为适应教育大数据中含有的大量类别信息,采用了 K-prototype 聚类方法对高校学生大数据进行聚类,以此构建学生画像[12]。许智宏等通过改进 K-means 算法和 PCA 算法来对学生行为进行用户画像[13]。凌玉龙等在引入马氏距离的基础上通过改变初始聚类中心的选择来改进 K-means 算法,从而适应学生群体聚类场景,更好地刻画学生的消费画像[14]。王惠惠等在实施学生群体画像的过程中为了提高聚类结果的精确性和鲁棒性,利用 KMeans、KModes 和 GMM 三种聚类方法构建基聚类器,并通过投票方法对聚类结果进行集成处理[15]。袁苗苗等基于改进的 K-means 聚类算法针对记录数据和用户评论数据分别建立了用户兴趣特征标签库和用户消费特征标签库,提出了多数据源融合的用户画像构建方法[16]。

由此可见,K-Means 聚类算法成为研究者构建用户画像时最常被采用的方法,但是 KMeans 等聚类算法鲁棒性不好,对噪声敏感,同时存在对离散型特征无法进行有效训练的缺陷。考虑到不同维度的特点,针对具有代表性的学习能力及学习行为标签,文中通过提出一种新的调整的线性加权变异系数算法,实现了学生学习能力标签模型;同时基于偏好随机变量概率分布理论,利用箱线图和 $k$ 百分位数方法构建了学生行为标签模型,较好地实现了学生画像的精准构建。

## 2　学习能力标签模型

文中使用的学生学习数据集按照教学周阶段性产生、采集,具有连续的数值型特征,同时也具备周期性、动态性的特点。对学习能力的阶段性刻画,集中趋势度指标是一种常用的方法,如均值、众数、中位数等,因为这些指标代表了学生的平均水平。但是均值的鲁棒性非常差,容易受到噪声的影响,而众数则更加适合离散的数据特征。中位数虽然克服了上述两种指标度量的缺点,兼具鲁棒性和数值特征适应性,但是又没有考虑到每一次成绩的变化波动情况。离中趋势度指标是另外一种可以用于刻画学习能力的方法。但是如果单纯使用方差或者标准差,虽然能够度量数据的离散程度,但是忽略了成绩数据的周期动态性特点,即每周都会有新的成绩数据产生。成绩数据集合以周为单位进行扩充,样本容量每周发生变化。因此采用变异系数(Coefficient of Variation,CV)的形式度量学习能力稳定性是较为合适的方法。CV 没有量纲,不受样本容量限制,这样就可以对学习能力稳定性进行客观比较。

传统的变异系数 CV 的计算方式为原始数据标准差与原始数据平均数的比,如式(1)所示:

$$CV(x)=\sigma_x/\bar{x}=\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}/\left(\sum_{i=1}^{n}x_i/n\right) \tag{1}$$

传统的变异系数 CV 计算方法简洁,但是没有考虑变量每一次取值的差异性与重要性,因此,该文引入了加权调整的变异系数 Adjusted_CV,解决带权重的特征稳定性的计算问题。

图 1 是构建学习能力稳定性的算法模型。

成绩数据源 SDataset 如式(2)所示,包括 $m$ 个学生,$n$ 次成绩。

$$SDataset=[s_1,s_2,\cdots,s_m]=\left[\begin{pmatrix} ws_{11} & ws_{12} & \cdots & ws_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ ws_{m1} & ws_{m2} & \cdots & ws_{mn} \end{pmatrix}\right] \tag{2}$$

其中,$S_i\{i=1,2,\cdots,m\}$ 为学生成绩样本,$ws_{i,j}$ 为第 $i$ 个样本第 $j$ 周的成绩( ws 为 weekscore 的简记),如式(3)所示:

$$s_i=\{ws_{i1},\ ws_{i2},\cdots,\ ws_{in}\}\overset{\text{def}}{=}\{weekscore_1,\cdots,weekscore_i,\cdots,weekscore_n\} \tag{3}$$

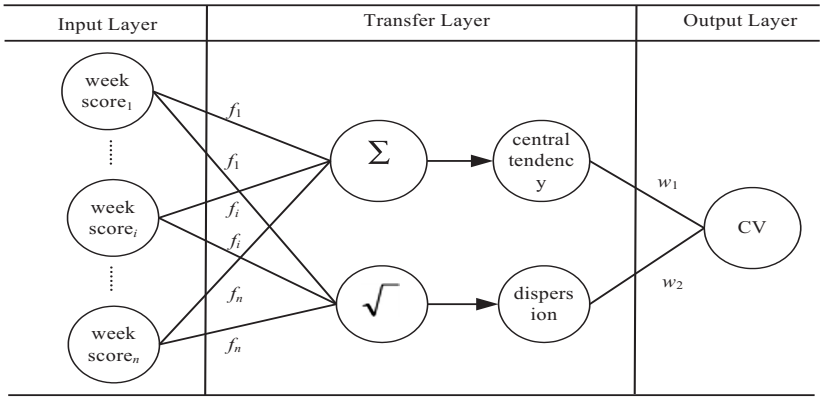


图 1 学习能力稳定性算法模型

可以通过图 1 所示的学习能力稳定性算法模型计算 $s_i$ 的 CV 系数值。模型输入层 InputLayer 接收到按周期采集的 $n$ 次成绩:$\text{weekscore}_1,\cdots,\text{weekscore}_n$,每次成绩根据其难度系数给予不同权重 $f_i$,$i$ 的取值为 1,2,…,$n$。转换层 TransferLayer 根据接收到的成绩及权重数据,计算集中趋势度和离中趋势度。集中趋势度采用加权线性平均的形式进行计算,计算结果记为 Weighted_Mean(score_stu),如式(4)所示:

$$\text{Weighted\_Mean}(\text{score\_stu}) = \sum_{i=1}^{n} f_i \times \text{weekscore}_i \tag{4}$$

其中,$f_i$ 为每次任务的难度系数权重,$i$ 的取值为 1,2,…,$n$。

离中趋势度的计算采用加权的样本标准差进行计算,其中 $n$ 为样本容量,即当前个体成绩数量。计算结果记为 Weighted_σ(score_stu),如式(5)所示:

$$\text{Weighted\_}\sigma(\text{score\_stu}) = \sqrt{\sum_{i=1}^{n} \frac{(f_i \times \text{weekscore}_i - \text{Weighted\_Mean}(\text{score\_stu}))^2}{(n-1)}} \tag{5}$$

其中,$\text{weekscore}_i$ 是动态的每周学习成绩,$n$ 为时间窗口期内的作业数量。

模型输出层 OutputLayer 计算最终的学习能力稳定性系数 CV 值,采用加权的标准差与加权线性均值的比值计算,进而调整的 Adjusted_CV 计算公式如式(6)所示:

$$\text{Adjusted\_CV}(\text{score}_i) = \frac{\text{Weighted\_}\sigma(\text{score\_stu})}{\text{Weighted\_Mean}(\text{score\_stu})} \tag{6}$$

其中,Adjusted_CV(score)作为个体成绩稳定性原始评价指标,可有效衡量窗口期内学生成绩的稳定性情况,消除量纲与样本容量的影响。Adjusted_CV(score)数值越小,窗口期内学生成绩越稳定地趋近于该学生的平均水平,集中趋势的代表性越好,学生的学习能力越稳定。Adjusted_CV(score)数值越大,平均成绩的代表性也就越差,成绩数值的震荡性越大,因而学生能力的稳定性也就越差。

经过上述算法对 Adjusted_CV 值的处理,可以得到一系列个体成绩稳定性原始评价数据集合。$\text{Adjusted\_CV}(\text{score}) = \{\text{score}_i, i = 1,2,\cdots,n\}$,$n$ 为样本容量。为了评价个体学生成绩稳定性在全量样本中的位置,此处采用箱线图 $k$ 百分位数的方式进行离散化,计算方法为 $p = 1 + (n-1) \times k\%$,$p$ 为 $k$ 百分位数的位置,此处 $k$ 的取值为序列 [0,25,50,75,100],从而最终产生个体学习稳定性标签。上述完整的学习能力稳定性标签构建算法如算法 1 所示。

算法 1:学习能力稳定性标签构建算法

输入:阶段性在线学习事务数据集 $C$

过程:

(1)Shuffle($C$) //随机打乱数据集

(2)For each score_stu in $C$:

(3) Aggregation(score_stu) //分组聚合个体样本的阶段性评分数据

(4) 根据式(4)计算 Weighted_Mean(score_stu) //计算个体线性加权集中趋势度指标

(5) 根据式(5)计算 Weighted_σ(score_stu) //计算个体加权离中趋势度指标

(6) 根据式(6)计算 Adjusted_cv($\text{score}_i$) //计算该个体成绩稳定性指标

(7) Add(CV, Adjusted_cv) //将个体成绩稳定性指标 Adjusted_cv 加入全量样本稳定性指标集合 CV

(8)End For

(9)Sort(CV) //对全量样本 cv 值进行排序

(10) $P = 1 + (n-1) \times k\%$ //计算箱线图 $k$ 百分位数,$P$ 为 $k$ 百分位数位置集合,$k$ 取值序列为 [0,25,50,75,100],$n$ 为样本数

(11)For each cv in CV:

(12) loc = Position(cv, $P$) //计算个体样本所处百分位数位置

(13) $F_i$ = AssignFlag(loc) //根据个体位置赋予对应标签

(14) Add($F$, $F_i$) //将个体成绩稳定性标签 $F_i$ 加入全

量样本稳定性标签集合 $F$

(15)End For

输出:学习成绩稳定性画像标签集合 $F$

## 3　学习行为标签模型

学习行为是指学生在线学习的行为习惯,如学习响应习惯、设备访问习惯、登录时间习惯、作业完成习惯等。其中学生对学习任务的响应习惯最具代表性,反映了学生的学习主动性和积极性。下面以学习响应习惯为例,详细阐述行为偏好类画像标签模型的构建算法。图 2 展示了学习响应习惯偏好行为的事务数据流。学习响应偏好数据的产生主要由任务点、作业、测试、讨论等行为触发,而终端个体会响应该任务,形成访问时间数据流。学习响应习惯偏好标签模型以全量时间数据流为基础,利用箱线图 $k$ 百分位点方法及概率分布等理论产生。相比较传统的忽略中间时刻敏感度、使用部分响应取平均的方式,这种构建方法更为精准客观。

第一步是单次行为事件的触发,将每一次任务的发布事件序列记为 $T=\{t_{\text{release}},t_{\text{check}},t_{\text{submit}}\}$。其中 $t_{\text{release}}$、$t_{\text{check}}$、$t_{\text{submit}}$ 分别为发布时间、查看时间和提交时间。切片时间段数据记为 $V=\{v_{\text{sensitive}},v_{\text{complete}}\}$,其中 $v_{\text{sensitive}}=t_{\text{check}}-t_{\text{release}}$,$v_{\text{complete}}=t_{\text{submit}}-t_{\text{check}}$。学习响应敏感度为任务查看时间减去任务发布时间,学习响应完成度为任务提交时间与查看时间之差。每一个个体一次任务的响应值计算公式如式(7)所示:

$$r_{j,i}=w_1*v_{\text{sensitive}}+w_2*v_{\text{complete}}$$
$$w_1+w_2=1,i=1,2,\cdots,m,j=1,2,\cdots,n \tag{7}$$

响应值 $r_{j,i}$ 即为响应敏感度和完成度的线性加权平均,$m$ 为发布任务数,$n$ 为学生样本量,$v_{\text{sensitive}}$ 为一次任务的学习响应敏感度,$v_{\text{complete}}$ 为一次任务的学习响应完成度,$w_1$、$w_2$ 分别为敏感度和完成度权重。

对于一次任务,全量学生形成的响应度集合为 $R_i=\{r_{1,i},r_{2,i},\cdots,r_{n,i}\}$。

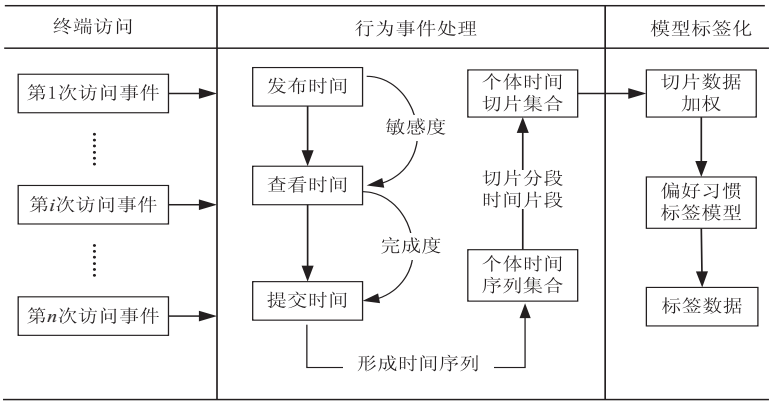


图 2　学习响应习惯偏好行为事务数据流

第二步,采用箱线图 $k$ 百分位数的方式对响应度集合 $R_i$ 进行离散化,计算方法为 $p=1+(n-1)\times k\%$,$p$ 为 $k$ 百分位数的位置,$k$ 的取值为序列[0,30,70,100]。

第三步,采用众数投票的方式对每一次任务分段结果进行投票计数,取分段频次最大概率值作为最终的学习响应习惯标签。分段概率计算公式如式(8)所示。

$$p_j=\begin{cases}p_{\text{pos}}=\text{n\_pos}/m,\text{n\_pos}=\text{n\_pos}+1,r_{j,i}\in[0,30\%]\text{分位点},i=1,2,\cdots,m\\ p_{\text{com}}=\text{n\_com}/m,\text{n\_com}=\text{n\_com}+1,r_{j,i}\in(30\%,70\%]\text{分位点},i=1,2,\cdots,m\\ p_{\text{neg}}=\text{n\_neg}/m,\text{n\_neg}=\text{n\_neg}+1,r_{j,i}\in(70\%,100\%]\text{分位点},i=1,2,\cdots,m\end{cases} \tag{8}$$

其中,n_pos、n_com、n_neg 为第 $j$ 个样本的积极性、一般、消极性的支持度计数,$m$ 为任务数,$p_j$ 为第 $j$ 个样本学习响应分段频次概率集合,$p_{\text{pos}}$ 为响应积极的概率,$p_{\text{com}}$ 为响应一般的概率,$p_{\text{neg}}$ 为响应消极的概率。最终的个体标签取决于概率分布的最大值,$\max P_j=\max\{p_{\text{pos}},p_{\text{com}},p_{\text{neg}}\}$。上述完整的学习响应习惯标签模型构建算法如算法 2 所示。

算法 2:学习响应习惯标签模型构建算法

输入:切片时间事件数据集 $C$

过程:

(1)For each $T_i$ in $C.T$:　//遍历学习任务数据集

(2)　For each $S_j$ in $T_i.S$:　//遍历第 $i$ 次任务的个体样本学习数据集

(3)　$S_j.v_{\text{sensitive}}=S_j.t_{\text{check}}-S_j.t_{\text{release}}$　//计算样本 $j$ 的任务敏感度

(4)　$S_j.v_{\text{complete}}=S_j.t_{\text{submit}}-S_j.t_{\text{check}}$　//计算样本 $j$ 的任务完成度

(5)　根据式(7)计算 $R_{j,i}$　//计算个体样本 $j$ 的第 $i$ 次任务的响应值

(6)　Add( $R_i$ , $R_{j,i}$ )　//将个体任务响应值 $R_{j,i}$ 加入全量样本响应值集合 $R$

(7) End For

(8) $P=1+(n-1)\times k\%$ //计算箱线图 $k$ 百分位数,$P$ 为 $k$ 百分位数的位置集合,$k$ 的取值为序列[0,30,70,100],$n$ 为个体样本数

(9) For each $R_{j,i}$ in $R_i$ :

(10) loc=Position( $R_{j,i}$ , $P$ ) //计算个体样本 $j$ 所处百分位数位置

(11) $MF_{j,i}$ =Flag(loc) //计算样本 $j$ 第 $i$ 次任务的标签

(12) Add(MF, $MF_{j,i}$ ) //将样本 $j$ 第 $i$ 次任务标签 $MF_{j,i}$ 加入全量样本任务积极性标签阶段性集合 MF

(13) End For

(14) End For

(15) For each $MF_j$ in MF:

(16) 根据式(8)计算 $P_j=\{P_{pos},P_{com},P_{neg}\}$ //计算个体学习响应分段频次概率集合

(17) $F_j=\max(P_j)$ //生成个体学习响应习惯标签,个体标签取决于概率分布的最大值

(18) Add( $F$ , $F_j$ ) //将个体响应习惯标签 $F_j$ 加入全量样本响应习惯标签集合 $F$

(19) End For

输出:学习响应习惯标签集合 $F$

## 4 实验结果与分析

实验数据通过学习通系统在线数据采集,并结合教务系统历史成绩等辅助信息进行人工标注。利用调整的线性加权变异系数算法进行学习能力稳定性模型实验,部分抽样数据及处理结果如表 1 所示。表中,$ws_i$ 表示周次,Linearwei_CV 表示调整后的 CV 值,Lw_CV_Quan 表示样本所处分位点,tra_tendency 表示样本成绩平均值。

表 1 调整的线性加权变异系数算法处理结果示例

| Stu_ID | $ws_1$ | $ws_2$ | $ws_3$ | $ws_4$ | $ws_5$ | $ws_6$ | $ws_7$ | $ws_8$ | $ws_9$ | $ws_{10}$ | $ws_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19＊＊＊＊＊04 | 94.5 | 100 | 100 | 100 | 100 | 95 | 100 | 100 | 100 | 100 | 100 |
| 19＊＊＊＊＊27 | 66 | 69 | 64 | 67 | 62 | 60 | 70 | 68 | 69 | 70 | 70 |
| 19＊＊＊＊＊02 | 94.5 | 90.6 | 80 | 90 | 100 | 100 | 92.5 | 98 | 99 | 100 | 94 |
| 19＊＊＊＊＊26 | 66 | 95.3 | 90 | 60 | 95 | 59 | 82.5 | 100 | 100 | 100 | 65 |
| 19＊＊＊＊＊07 | 71.5 | 84.6 | 100 | 55 | 49 | 76 | 40 | 100 | 100 | 100 | 90 |
| 19＊＊＊＊＊28 | 61.5 | 81.2 | 30 | 95 | 71 | 74 | 45 | 100 | 70 | 80 | 30 |

| Stu_ID | $ws_{12}$ | $ws_{13}$ | $ws_{14}$ | $ws_{15}$ | Linearwei_CV | Lw_CV_Quan | tra_tendency |
|---|---|---|---|---|---|---|---|
| 19＊＊＊＊＊04 | 95.3 | 100 | 95.5 | 100 | 0.02 | Q1 | 98.69 |
| 19＊＊＊＊＊27 | 67 | 70 | 65 | 70 | 0.05 | Q1 | 67.13 |
| 19＊＊＊＊＊02 | 90 | 91 | 90 | 89 | 0.16 | Q2 | 93.24 |
| 19＊＊＊＊＊26 | 100 | 100 | 86.5 | 100 | 0.23 | Q3 | 86.62 |
| 19＊＊＊＊＊07 | 100 | 95 | 82 | 100 | 0.25 | Q3 | 82.87 |
| 19＊＊＊＊＊28 | 75.2 | 95 | 50.5 | 100 | 0.45 | Q4 | 70.56 |

从表 1 可以看出,序号为 19＊＊＊＊＊04、19＊＊＊＊＊27 的两个样本在文中所采用的变异系数方法中系数值分别为 0.02、0.05,在全量样本中位于第 Q1 分位点处,成绩稳定性都很高,4 号样本成绩高且稳定在 98.69 附近,27 号样本成绩低且稳定在均值 67.13 附近。19＊＊＊＊＊02 在全量样本中位于第 Q2 分位点处,成绩稳定性良好,在均值附近有一定的波动,但与均值的偏差不大。19＊＊＊＊＊26、19＊＊＊＊＊07,在全量样本中位于第 Q3 分位点处,成绩稳定性一般,震荡较明显。19＊＊＊＊＊28 在全量样本中位于第 Q4 分位点处,成绩稳定性差,各次成绩与平均值 70.56 的偏差较大,震荡明显。

利用调整的线性加权变异系数 Adjusted_CV 算法与传统的变异系数算法进行学习能力稳定性对比实验,模型效果如图 3 所示。相较于传统的变异系数算法,调整权重后的 Adjusted_CV 算法具有更好的拟合效果。

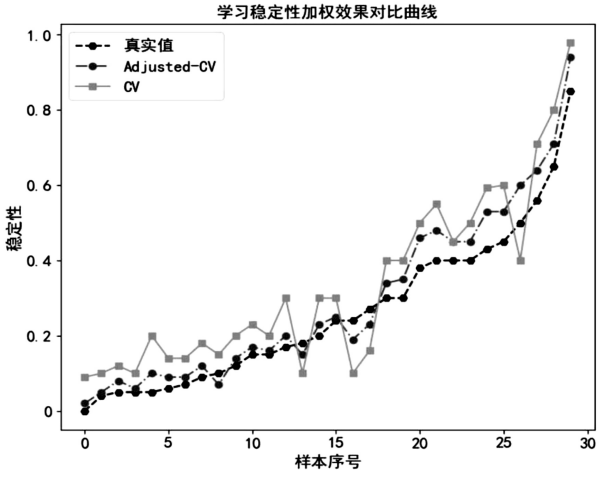


图 3 学生学习稳定性加权效果对比曲线

通过学习通系统累计采集 16 周的在线学习行为数据并进行人工标注,利用箱线图 $k$ 百分位数及随机

变量概率分布的组合方法进行学习响应习惯标签模型实验,部分抽样数据及处理结果如表 2 所示。表中,Ti_release 表示第 $i$ 次任务的发布时间,Ti_check 表示第 $i$ 次任务的查看时间,Ti_submit 表示第 $i$ 次任务的提交时间,sensitive 表示敏感度,complete 表示完成度,vote 表示样本第 $i$ 次任务的标签,$P$(pos)表示样本积极性概率,$P$(com)表示样本一般性概率,$P$(neg)表示样本消极性概率,total 表示样本响应习惯最终标签。

表 2　箱线图 $k$ 百分位数及随机变量概率分布方法处理结果示例

| Stu_ID | T1_release | T1_check | T1_submit | sensitive | complete | vote | … | T16_release |
|---|---|---|---|---|---|---|---|---|
| 19＊＊＊＊＊02 | 2020-09-01 22:01:00 | 2020-09-02 10:31:55 | 2020-09-02 11:16:17 | 13 | 1 | pos | … | 2020-12-9 08:44:24 |
| 19＊＊＊＊＊04 | 2020-09-01 22:01:00 | 2020-09-02 09:08:28 | 2020-09-03 21:06:33 | 11 | 24 | com | … | 2020-12-9 8:44:24 |
| 19＊＊＊＊＊07 | 2020-09-01 22:01:00 | 2020-09-05 09:21:52 | 2020-09-05 09:39:30 | 12 | 1 | pos | … | 2020-12-9 08:44:24 |
| 19＊＊＊＊＊26 | 2020-09-01 22:01:00 | 2020-09-06 09:54:56 | 2020-09-06 10:08:44 | 12 | 1 | pos | … | 2020-12-9 08:44:24 |
| 19＊＊＊＊＊27 | 2020-09-01 22:01:00 | 2020-09-01 22:01:56 | 2020-09-06 12:15:43 | 0 | 123 | neg | … | 2020-12-9 08:44:24 |
| 19＊＊＊＊＊28 | 2020-09-01 22:01:00 | 2020-09-02 10:37:30 | 2020-09-02 11:05:36 | 13 | 1 | pos | … | 2020-12-9 08:44:24 |

| Stu_ID | T16_check | T16_submit | sensitive | complete | vote | $P$(pos) | $P$(com) | $P$(neg) | total |
|---|---|---|---|---|---|---|---|---|---|
| 19＊＊＊＊＊02 | 2020-12-18 15:10:24 | 2020-12-19 16:51:57 | 247 | 25 | pos | 0.88 | 0.12 | 0 | pos |
| 19＊＊＊＊＊04 | 2020-12-09 12:28:56 | 2020-12-10 22:56:50 | 4 | 34 | com | 0.75 | 0.25 | 0 | pos |
| 19＊＊＊＊＊07 | 2020-12-13 11:41:44 | 2020-12-19 12:02:17 | 89 | 144 | Neg | 0.06 | 0.44 | 0.50 | nes |
| 19＊＊＊＊＊26 | 2020-12-09 15:29:21 | 2020-12-10 18:12:53 | 7 | 27 | pos | 0.75 | 0.19 | 0.06 | pos |
| 19＊＊＊＊＊27 | 2020-12-16 08:23:56 | 2020-12-19 20:41:18 | 168 | 84 | com | 0.13 | 0.63 | 0.25 | com |
| 19＊＊＊＊＊28 | 2020-12-10 16:16:45 | 2020-12-13 14:26:53 | 32 | 23 | pos | 0.75 | 0.06 | 0.19 | pos |

从表 2 可以看出,19＊＊＊＊＊02、19＊＊＊＊＊04、19＊＊＊＊＊26、19＊＊＊＊＊28 四个样本对历次任务响应比较积极,其中 19＊＊＊＊＊02 积极响应的占比达 88%。从上述样本的过程细节数据来看,积极响应的个体样本历次任务的完成度较为及时。19＊＊＊＊＊27、19＊＊＊＊＊07 号样本响应程度分别为一般和消极,占比分别为 63%、50%。从这些样本的过程细节数据来看,此类样本单次任务响应敏感度和完成度较差,尤其是 19＊＊＊＊＊07 号样本虽然有时查看任务及时,但是执行力很差,有严重的拖沓习惯。

通过基于箱线图 $k$ 百分位数及随机变量概率分布的方法可以得出学生响应偏好识别结果的混淆矩阵,如图 4 所示。从图中可知,方法的准确率为 83%,识别效果良好,能够很好地刻画个体的响应习惯偏好。

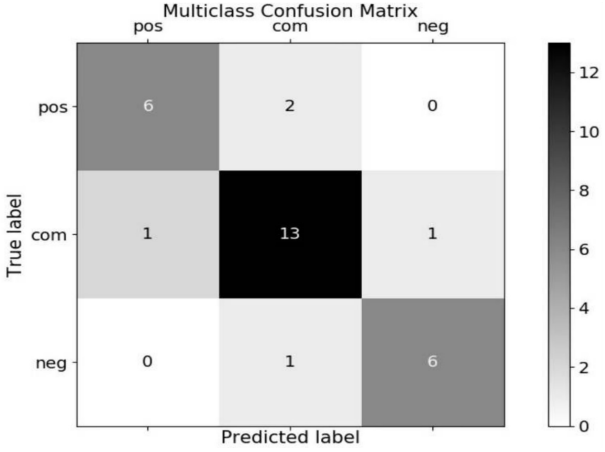


图 4　学生响应偏好识别结果混淆矩阵

## 5　结束语

混合式学习积累了海量的学生学习数据。充分挖

掘和利用这些学习过程和学习结果数据,实施学生学习画像是面向未来型教育的一个重要研究领域。学习画像能够很好地刻画学生在学习能力、学习行为和学习成效等方面的特征,实现学生群体的划分[17-18],通过数据驱动更好地为个性化学习规划学习路径[19-20]。学习画像的关键在于对学生学习各个特征维度的标签模型进行构建,从数据的分析结果中提炼出合适的标签来对目标对象的学习特征进行标识。文中提出的一种调整的线性加权变异系数算法,以及对偏好随机变量概率分布理论和箱线图 $k$ 百分位数方法的应用,成功地构建了学习画像中最关键的学习能力和学习行为两个维度的标签模型。实验结果的对比分析也证明了这种构建方法的合理性和有效性。在后续模型优化过程中,可以考虑扩充数据维度、调整过程权重等方式进一步优化模型效果。

**参考文献:**


[1] 赵雅慧,刘芳霖,罗 琳. 大数据背景下的用户画像研究综述:知识体系与研究展望[J]. 图书馆学研究,2019(24):13-24.

[2] 徐 芳,应洁茹. 国内外用户画像研究综述[J]. 图书馆学研究,2020(12):7-16.

[3] GOEL S,KUMAR R. Folksonomy-based user profile enrichment using clustering and community recommended tags in multiple levels[J]. Neurocomputing,2018,315(13):425-438.

[4] DESPENIC M,CHRAIBI S,LASHINA T. Lighting preference profiles of users in an open office environment[J]. Building and Environment,2017,116(3):89-107.

[5] VIEIRA C,PARSONS P,BYRD V. Visual learning analytics of educational data:a systematic literature review and research agenda[J]. Computers & Education,2018,122:119-135.

[6] 余明华,张 治,祝智庭. 基于可视化学习分析的研究性学习学生画像构建研究[J]. 中国电化教育,2020(12):36-43.

[7] 杨长春,徐 筱,宦 娟,等. 基于随机森林的学生画像特征选择方法[J]. 计算机工程与设计,2019,40(10):2827-2834.

[8] 黄文林. 基于学生画像分析的高校精准思政探索[J]. 东北大学学报:社会科学版,2021,23(3):104-111.

[9] 任红杰. 基于大数据的精准教学:生成路径与实现条件[J]. 黑龙江高教研究,2017(9):165-168.

[10] 杨彩霖. 基于大数据的师生教学行为画像的构建与应用研究[J]. 高教学刊,2021,7(31):78-81.

[11] 张 毅. 大数据背景下用户画像的统计方法实践分析[J]. 现代商业,2020(6):9-10.

[12] 翟鸣宇,程 建,王苏桐,等. 基于 K-prototype 聚类的学生教育画像分析[J]. 大连理工大学学报:社会科学版,2021,42(6):22-31.

[13] 许智宏,李彤彤,董永峰,等. 基于改进 K-means 算法的学生用户画像构建研究[J]. 河北工业大学学报,2022,51(3):19-24.

[14] 凌玉龙,张 晓,李 霞,等. 改进 Kmeans 算法在学生消费画像中的应用[J]. 计算机技术与发展,2021,31(10):122-127.

[15] 王惠惠,董永权,和文斌,等. 基于聚类集成的学生群体画像构建[J]. 江苏师范大学学报:自然科学版,2022,40(3):46-50.

[16] 袁苗苗,侯瑞春,陶 冶,等. 基于多数据源融合的用户画像构建方法[J]. 计算机与数字工程,2022,50(4):757-761.

[17] DINH D P,HARADA F,SHIMAKAWA H. Directing all learners to course goal with enforcement of discipline utilizing persona motivation[J]. IEICE Transactions on Information and Systems,2013,96(6):1332-1343.

[18] KHALIL M,EBNER M. Clustering patterns of engagement in massive open online courses (MOOCs):the use of learning analytics to reveal student categories[J]. Journal of Computing in Higher Education,2017,29(1):114-132.

[19] SMET C D,SCHELLENS T,WEVER B,et al. The design and implementation of learning paths in a learning management system[J]. Interactive Learning Environments,2016,24(6):1076-1096.

[20] 师亚飞,彭红超,童名文. 基于学习画像的精准个性化学习路径生成性推荐策略研究[J]. 中国电化教育,2019(5):84-91.