基于数据挖掘的开放教育学习者学习行为聚类分析

杨孟娇，侯丽媛

（内蒙古广播电视大学，呼和浩特，010011）

摘要：基于学习者在国家开放大学学习网的海量在线学习行为数据的分析、处理，根据研究对象和研究内容选取了 7 个学习行为量化指标，采用主成分分析法和 K 均值聚类算法相结合的方法，对开放教育学习者的学习行为进行科学的聚类，提取了行为数、浏览数、形考完成率和形考及格率四个主成分因子，分别以四个指标进行聚类并给出了四个评价等级的聚类结果。研究结果表明，开放教育学习者的学习行为数、浏览数水平整体偏低，但是形考作业的完成率和及格率都达到了较高的水平。通过分析聚类结果，可以准确识别出在线行为数和浏览数不高、形考作业完成率低、形考作业及格率偏低的学习者，针对各类学习者可以针对性地采取相应的措施，从而提升学习效果。

关键词：教育大数据；在线学习行为；因子分析；K 均值聚类算法

中图分类号：G724.82；G434　　**文献标识码**：A　　**文章编号**：1008-6021(2021)03-0032-06

开放教育为学习者自主学习提供了海量学习资源，但是因为学习者面临工学矛盾以及本身的主观惰性，使得不同学习者的学习行为存在差异。因此掌握学习者的学习行为现状、学习特征，是提升开放教育教学质量、为学习者提供更好的学习支持服务的基础与前提。对学习者在线学习行为的研究已经是热点与重点。洪宣容[1]等人从学习者特征、在线学习行为、在线学习资源、现有学习支持服务、在线学习评价等方面对成人在线学习进行了现状研究。薛瑞璇[2]将学习行为分为外显行为和内隐行为两种，外显行为主要有学习者登录、浏览、发帖提问、参与讨论等具体操作行为；内隐行为主要有初始能力分析、学习风格分析、学习动机分析和自我效能感分析等。舒忠梅[3]等人基于数据挖掘，将学习行为分为全面发展型、均衡发展型等七个类型。丁鹏飞[4]基于学生学习特征将学生聚类划分成分别命名为被动型、游离型、徒劳无功型、学有余力型，并对不同类型学生的学习投入模式及其行为特征进行分析。杨国龙[5]构建了网络教育学习者学习行为差异化分析模型，聚类并分析了内向场独立、内向场依存、外向场独立与外向场依存四类学习者群体的特征差异。王红梅[6]等对 408 名学习者的学习行为数据进行分析，探究开放学习环境中学习行为投入与认知投入的关系。田娜[7]以“程序设计语言 C”的 65 个学生为样本，对学生完成课时数量、浏览次数、登录次数等指标进行聚类分析。王蕾[8]对开放教育学习者的登录、浏览、交互、检索等各类数据进行分析研究，就如何提高学生参与在线学习的积极性提出改进路径和建议。

作为开放教育的主阵地，国家开放大学学习网是国家开放大学以及全国各分部的主要学习平台，开放教育学习者的学习行为主要也集中在此。本文充分利用国家开放大学学习网的数据，以内蒙古分部的学习者为研究对象，提取学习者学习行为的特征参数，运用主成分分析和 K 值聚类算法，探究开放教育学习者学习行为特征与规律，筛选出主要指标以及评估结果偏低的学习者，针对性提出有效措施，以期为学习者提供更好的学习支持服务、教学服务。

一、数据采集与指标确定

（一）数据采集与处理

本文的数据来源于国家开放大学学习网，选取内

收稿日期：2021-04-08

基金项目：内蒙古自治区高等学校科学研究项目“基于在线教学数据分析的网络核心课程教学实施团队建设研究——以内蒙古为例”（项目编号：NJSY20290）；内蒙古广播电视大学“活学活用计算机”教师协同创新工作室项目

作者简介：杨孟娇（1990—），女，内蒙古呼和浩特人，讲师，硕士。研究方向：远程教育技术。

蒙古分部2018年秋到2020年春四个学期的32万多条学习者学习行为数据，原始数据表包括了22个指标(表1)。由于国开学习平台上获取的为原始数据，包括了所有选课以及未选课的学习者学习行为数据，故研究删除了选课数为零的无效数据；对于原始数据存在个别变量缺失的问题，通过将缺失值替换为变量的平均值进行缺失数据处理，经过处理最终得到有效数据186 497条。

表1　指标名称

序号	1	2	3	4	5	6	7	8	9	10	11
指标名称	分部编号	分部名称	学院编号	学院名称	学习中心编号	学习中心名称	姓名	行政班级名称	发生行为省校拼音	发生行为省校	选课数
序号	12	13	14	15	16	17	18	19	20	21	22
指标名称	学习平台登录次数	学习平台在线天数	行为总数	浏览数	完成形考课程数	形考及格课程数	提交形考数	形考评阅数	发帖数	回帖数	教师回帖数

(二)指标确定

研究对象设定为学习者，所以表1中指标1到10均为无关变量，学习者姓名替换为序号即可；研究的内容为学习者学习行为，因此只关注学习者学习平台的相关行为，不考虑教师的相关行为，包括形考评阅数和教师回帖数；提交形考数与完成形考课程数存在较大关联，而且完成形考更为重要，所以也不考虑提交形考数。因此本文重点研究的是反映学习者学习积极性、完成度、参与度等特征的多项指标，并基于原始数据确定如下指标。

1.学习平台登录次数、学习平台在线天数

学习平台登录次数是指一个学期内学习者登录国开学习网的次数，学习平台在线天数是指一个学期内学习者国开学习网的在线天数。由于不同学期、不同学习者的选课数量不同，在此将学习平台登录次数设定为平均每门课程的登录次数，将学习平台在线天数设定为平均每门课程的在线天数。

$$LN = LN_t / n \tag{1}$$

$$ON = ON_t / n \tag{2}$$

式中：LN 表示学习平台平均登录次数；LN_t 是原始数据中学习平台登录次数，也就是总次数；n 是每个学习者选课数量；ON 是学习平台平均在线天数；ON_t 是原始数据中学习平台在线天数，也就是总天数。

2.行为数、浏览数

行为数是指一个学期内学习者在国开学习网学习的所有行为数，包括浏览学习资源数、发帖数、回帖数、完成作业等行为数；浏览数是指一个学期内学习者浏览学习文本资源、视频资源等的数量。在此将行为数设定为平均每门课程的行为数，将浏览数设定为平均每门课程的浏览数。

$$B = B_t / n \tag{3}$$

$$R = R_t / n \tag{4}$$

式中：B 表示行为数；B_t 是原始数据中行为总数；R 是浏览数；B_t 是原始数据中的浏览数，也就是浏览总数。

3.形考完成率、形考及格率

形考完成率是指一个学期内学习者在学习平台完成形考课程数与选课数的比值；形考及格率是指一个学期内学习者在学习平台形考及格课程数与完成形考课程数的比值。

$$C = C_n / n \tag{5}$$

$$P = P_n / C_n \tag{6}$$

式中：C 表示形考完成率；C_n 是原始数据中完成形考课程数；P 是形考及格率；P_n 是原始数据中形考及格课程数。

4.发帖数、回帖数

发帖数是指一个学期内学习者在学习平台网上教学环节、讨论区等发布帖子数量；回帖数是指一个学期内学习者在学习平台网上教学环节、讨论区等回复他人帖子的数量。在此将发帖数设定为平均每门课程的发帖数，将回帖数设定为平均每门课程的回帖数。

$$PN = PN_t / n \tag{7}$$

$$RN = RN_t / n \tag{8}$$

式中：PN 表示发帖数；PN_t 是原始数据中的发帖

数，也就是发帖总数；RN 是回帖数；RN_t 是原始数据中的回帖数，也就是回帖总数。

通过公式(1)到(8)分别对原始数据进行处理，得到本文研究样本数据。

二、因子分析

因子分析是指研究从变量群中提取共性因子的方法，它是在尽可能降低原有信息损失的前提下，将本质相同的变量归入一个因子，可以减少变量的数量，也就是指标降维[9]。本文选择主成分分析法进行因子分析。

(一)可行性分析

采用 KMO 法和巴特利特法对数据的可行性进行检验[10]。表 2 中，KMO 值为 0.774，大于 0.6，表示样本数据各指标之间具有较好的相关性，适用于因子分析。巴特利特球形度检验显著性水平为 0.000，表示拒绝零假设，样本数据变量之间存在相关性，适合进行主成分分析。

表 2　KMO 和巴特利特检验结果

KMO 取样适切性量数		0.774
巴特利特球形度检验	近似卡方	1 022 843.626
	自由度	28
	显著性	0.000

根据公因子方差可以判断公因子对各个指标的说明程度，提取公因子方差越大，公因子对对应指标的说明程度越大，公因子方法越小表示该指标的重要度越低，一般指标的提取公因子方差小于 0.4，就可以认为重要度较低，可以在因子分析中删除，因此由表 3 可以判断将回帖数删除不予分析。重新进行 KMO(0.771)和巴特利特检验(0.000)，结果均可行。

表 3　公因子方差

	初始	提取
学习平台登录次数	1.000	0.723
学习平台在线天数	1.000	0.603
行为数	1.000	0.887
浏览数	1.000	0.906
形考完成率	1.000	0.670
形考及格率	1.000	0.656
发帖数	1.000	0.514
回帖数	1.000	0.090

(二)提取主因子

可以通过碎石图确定最优主因子的数量，横坐标的组件号即为因子的数量，纵坐标是因子特征值，将因子特征值连线，较为陡峭的部分就是应该提取的主因子数量。由图 1 可知，前面两个因子的特征值较大且连线较陡，可以确定为主因子。

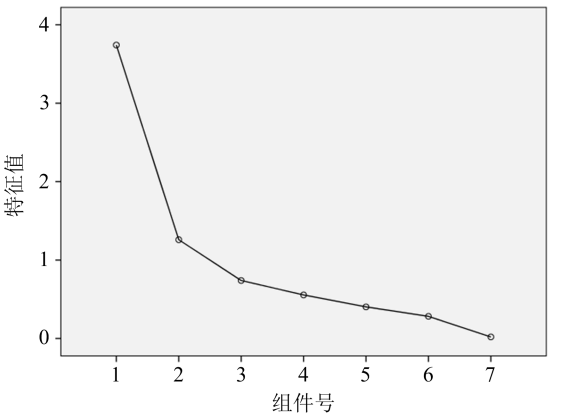

图 1　碎石图

主成分分析的目的是指标降维，但是同时也要尽可能地减少信息损失。因此在碎石图确定主因子数量的基础上，要求累积方差贡献率达到 85%以上，各成分在因子分析中的总方差解释结果如表 4 所示。由表 4 可知，两个主因子的累积方差贡献率为 71.445%，未达到 85%。由于成分 3、4 的特征值大于 0.5，因此可以将成分 3、4 也确定为主因子，累积方差贡献率达到 89.936%，符合要求。

表 4　总方差解释

成分	初始特征值		
	总计	方差百分比	累积 %
1	3.742	53.454	53.454
2	1.259	17.990	71.445
3	0.739	10.564	82.008
4	0.555	7.928	89.936
5	0.403	5.753	95.689
6	0.282	4.031	99.720
7	0.020	0.280	100.000

分析成分矩阵(表 5)，可以得知：主因子 1 与浏览数、行为数相关性较强，与其他指标均存在不同程度的相关性；主因子 2 与形考及格率和形考完成率有较强的相关性，与学习平台在线天数也存在弱相关。

表 5　成分矩阵

	成分	
	1	2
浏览数	0.944	−0.135
行为数	0.935	−0.127
学习平台登录次数	0.841	−0.155
学习平台在线天数	0.763	0.148
发帖数	0.638	−0.328
形考及格率	0.190	0.802
形考完成率	0.493	0.654

三、基于聚类的学生学习行为特征分析

（一）K 均值聚类算法

K 均值聚类算法（K-means 聚类算法）是集简单和经典于一身的基于距离的聚类算法。它采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大[11]。它是一种迭代求解的聚类分析算法，首先确定 k 值，即数据集经过聚类得到 k 个集合，也就是分类数，随机选取 k 个对象作为初始的聚类中心（质心），然后计算每个数据与各个聚类中心的距离，离哪个聚类中心近，就划分到那个聚类中心所属的集合。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算，直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

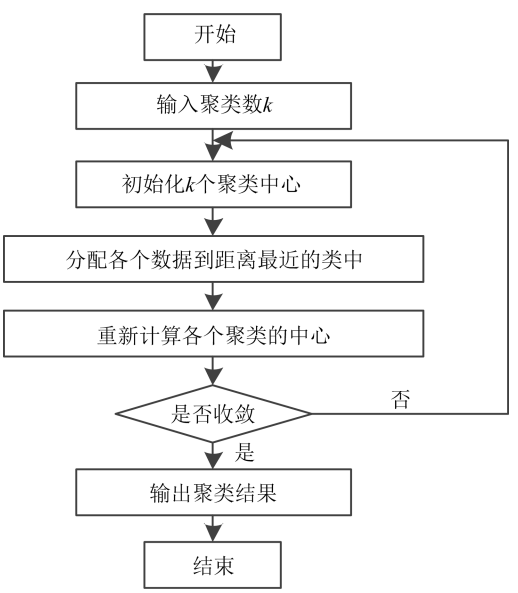

图 2　K 均值聚类算法流程

（二）学习行为特征聚类分析

根据因子分析的结果，分别对行为数、浏览数、形考完成率和形考及格率进行聚类，分析学习者的学习行为特征。将学习行为评价结果分为 4 类：高、较高、一般、偏低。图 3 将四个指标的评价结果相同的一类用线相连接，可以看出行为数、浏览数的聚类结果为偏低的占比较高，形考及格率的聚类结果为高的占比较高。具体聚类结果见表 6 到表 9。

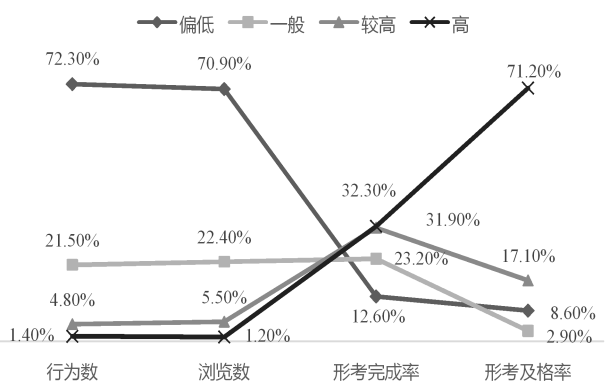

图 3　不同指标下各聚类结果占比

以行为数作为指标的聚类结果（表 6），平均每门课程的行为数在 0 到 159 次的学习者占比最高，达到 72.3%，评价结果偏低；行为数在 159 到 470 次的学习者占比为 21.5%，评价结果一般；仅有 6.2%的学习者的行为数在 470 以上，评价结果为较高或高。

表 6　以行为数作为指标的聚类结果

聚类	每个聚类中的个案数目及比例		最终聚类中心	聚类结果范围	评价结果
1	134 653	72.3%	57.51	0～159	偏低
2	40 204	21.5%	260.49	159.06～470.78	一般
3	8 970	4.8%	681.15	470.89～1010.86	较高
4	2 670	1.4%	1 340.64	1 011－6 154	高

以浏览数作为指标的聚类结果(表 7),平均每门课程的浏览数在 0 到 87 次的学习者占比最高,达到 70.9%,评价结果偏低;浏览数在 87 到 265 次的学习者占比为 22.4%,评价结果一般;有 5.5%的学习者的浏览数在 265 到 600 次之间,评价结果为较高;仅有 1.2%的学习者的浏览数在 600 以上,评价结果为高。

表 7　以浏览数作为指标的聚类结果

聚类	每个聚类中的个案数目及比例		最终聚类中心	聚类结果范围	评价结果
1	132 240	70.9%	31.39	0～87.67	偏低
2	41 762	22.4%	144.00	87.70～265.20	一般
3	10 367	5.5%	386.47	265.25～600.80	较高
4	2 128	1.2%	815.28	601.00～5 969.50	高

以形考完成率作为指标的聚类结果(表 8),整体评价结果偏高,每门课程的形考完成率在 85%到 100%之间的学习者占比最高,达到 32.3%,评价结果为高;形考完成率在 57%到 83%之间的学习者占比为 31.9%,评价结果较高;但是也有 12.6%的学习者的形考完成率在 28%以下。

表 8　以形考完成率作为指标的聚类结果

聚类	每个聚类中的个案数目及比例		最终聚类中心	聚类结果范围	评价结果
1	23 432	12.6%	0.13	0～0.28	偏低
2	43 239	23.2%	0.43	0.29～0.57	一般
3	59 582	31.9%	0.71	0.57～0.83	较高
4	60 244	32.3%	0.97	0.85～1	高

以形考及格率作为指标的聚类结果(表 9),整体评价结果很高,形考及格率在 50%以上的达到 88.3%,每门课程的形考及格率在 85%到 100%之间的学习者占比最高,达到 71.2%,评价结果为高;形考及格率在 50%到 83%之间的学习者占比为 17.1%,评价结果较高;仅有 8.6%的学习者的形考及格率在 15%以下。

表 9　以形考及格率作为指标的聚类结果

聚类	每个聚类中的个案数目及比例		最终聚类中心	聚类结果范围	评价结果
1	16 156	8.6%	0.00	0～0.15	偏低
2	5 550	2.9%	0.31	0.17～0.46	一般
3	31 921	17.1%	0.68	0.5～0.83	较高
4	132 870	71.2%	1.00	0.85～1	高

四、结论

基于对数据进行分析、挖掘,应用主成分分析和 K 均值聚类分析方法,对开放教育学习者的学习行为进行了科学聚类分析,其学习行为现状如下:学习者的学习行为数、浏览数水平整体偏低,但是形考作业整体情况较好,无论是完成率还是及格率都达到了较高的水平。通过分析可以对国开学习网在线行为数和浏览数不高、形考作业完成率低、形考作业及格率偏低的学习者进行准确地识别,针对这类学习者可以采取定时提醒督促学习和完成作业、提供针对性辅导等措施,提高其学习积极性以及学习的有效性。基于本文的研究内容及研究结论,通过国开学习网、学生教务和考务系统等提取更多有效的、有价值的学习行为特征指标,结合学习者自身属性特征、学习支持服务等,可以建立更加科学、完善的开放教育学习者学习行为评价体系,对开放教育管理人员、教师、学习支持服务人员以及学习者本身均有积极意义。

参考文献：

[1] 洪宣容，洪涛. 成人在线学习的研究现状与发展趋势[J]. 成人教育，2019，39(2)：25-29.

[2] 薛瑞璇. 在线学习平台中学习者的网络学习行为分析：以云南省工业人才在线学习网为例[D]. 昆明：云南大学，2016：19-25.

[3] 舒忠梅，徐晓东，屈琼斐. 基于数据挖掘的学生投入模型与学习分析[J]. 远程教育杂志，2015，33(1)：39-47.

[4] 丁鹏飞. 学习分析技术在教学中的应用研究[J]. 实验室研究与探索，2019，38(4)：215-219.

[5] 杨国龙. 网络教育学习者学习行为差异化研究[D]. 北京：北京邮电大学，2019：30-48.

[6] 王红梅，张琪，黄志南. 开放学习环境中学习行为投入与认知投入的实证研究[J]. 现代教育技术，2019，29(12)：48-54.

[7] 田娜，陈明选. 网络教学平台学生学习行为聚类分析[J]. 中国远程教育，2014(11)：38-41.

[8] 王蕾. 高等远程教育学生在线学习行为现状及改进路径研究：以国家开放大学网络课程学习为例[J]. 吉林广播电视大学学报，2019(7)：17-18.

[9] 陈碎雷，潘中柱. 高职院校现代学徒制试点：影响因子分析及改进策略[J]. 职业技术教育，2019，40(12)：18-22.

[10] 邵珠贵. 远程开放教育学生自主学习能力评价的研究[J]. 职业技术教育，2014，35(14)：38-41.

[11] 刘菲菲. 高职院校混合式教学实施状况及影响因素分析：基于X校网络教学综合平台的数据分析[J]. 职业技术教育，2019，40(26)：43-47.

Cluster Analysis of Learners' Learning Behavior in Open Education Based on Data Mining

YANG Mengjiao, HOU Liyuan

(Inner Mongolia Radio and Television University, Hohhot 010011, China)

Abstract: Based on the analysis and processing of massive online learning behavior data of learners in the National Open University Learning Network, seven quantitative indicators of learning behavior are selected according to the research object and research content, and the method of combining principal component analysis and K-means clustering algorithm is adopted to scientifically cluster the learning behaviors of open education learners, and extract four principal component factors: the number of behaviors, the number of views, the completion rate of the formative evaluation and the passing rate of formative evaluation, which are clustered by four indicators and given the clustering results of the four evaluation levels. The research results show that the number of learning behaviors and browses of open education learners are relatively low, but the completion rate and passing rate of the formative assessment homework have reached a high level. By analyzing the clustering results, the learners with low online behaviors and browses, low completion rate and passing rate of formative evaluation can be accurately identified. Relevant measures can be taken for all kinds of learners to improve the learning effect.

Keywords: big data in education; online learning behavior; factor analysis; K-means clustering algorithm

[责任编辑　许炎]