

基于线性回归算法的在线学习行为分析

郭玲玲, 范思萌, 王 梅, 苏冬娜

(东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318)

摘 要: 互联网的快速发展带动了教育领域的发展, 促使在线学习迅速兴起并深受教育人士认可。因此各种在线学习平台中的教学数据飞速递增, 对于如何充分利用、深度分析平台存储数据的价值, 引起了教育从事人员的关注。应用机器学习技术, 对学习者在在线学习行为与学习结果之间的相关性进行了一系列学习分析。通过收集在线平台上学生学习时的数据, 对收集到的数据进行预处理。基于 K-means 聚类算法对学习者的聚类建模, 将学生聚类成不同的类型。教师给予不同群体的学生相应的资料, 有效地提高学生的学习效率。基于线性回归算法分析学生学习行为, 确定学习行为对于学生最终成绩影响程度, 教师在众多在线平台学生活动中筛选出对学生最终成绩影响较大的活动, 完善教学方式。

关键词: 在线学习行为; K-means 聚类算法; 线性回归; 学习行为分析; 以学习者为中心

中图分类号: TP311.13; G434

文献标识码: A

文章编号: 1673-629X(2022)07-0191-05

doi: 10.3969/j.issn.1673-629X.2022.07.033

Online Learning Behavior Analysis Based on Linear Regression Algorithm

GUO Ling-ling, FAN Si-meng, WANG Mei, SU Dong-na

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: The rapid development of the Internet has driven the development of the education field, prompting the rapid rise of online learning and recognized by educators. Therefore, the teaching data from various online learning platforms increase rapidly. How to fully utilize and deeply analyze the value of the data stored on the platform has aroused the attention of educators. A series of learning and analysis on the correlation between learners' online learning behavior and learning outcomes were conducted by applying machine learning technology. Firstly, the studying data of students from online platforms were collected and preprocessed. Secondly, the K-means clustering algorithm was used to cluster the learners into different types. Teachers give the students in different groups the appropriate material, improving their learning efficiency effectively. Finally, the students' learning behavior was analyzed by the linear regression algorithm to determine the influence degree of learning behavior on students' final scores. From the numerous online activities, teachers can screen out the activities that have a great influence on students' final scores and improve the teaching methods.

Key words: online learning behavior; K-means clustering algorithm; linear regression; learning behavior analysis; learner centered

0 引 言

网络教育是通过互联网途径传播教学内容的一种方法。在“网络+教育”的推动下,网络教育得到了迅速发展,各种教育资源在该平台上不断完善。这些平台上汇聚了大量的资源,学习者和教育从业人士,形成了平台的大数据^[1]。这些大数据由学习者和教育人士在教学过程中的各种动态数据组成。这就促使在线分析成为教育领域的研究重点之一,它根据获取到的学生在线平台产生的动态数据^[2],利用数据分析算法对学生实际的学习行为进行分析,根据分析结果达到提

升教师教学水平与学生学习能力的目的^[3]。

选取该校计算机科学与技术专业学生及相关课程为研究对象,从在线学习平台的学生学习行为分析着手,跟踪课程学习过程,利用机器语言中的算法对学生的行为进行分析,以便从中挖掘学习行为与学习结果之间的映射关系,从而使得教师能够更客观更科学地了解到影响学生成绩的因素^[4]。教师还可以根据每个学生的学习行为来分析学生学习特点,采取针对性教学,根据不同的情况给予不同的建议,有效改善学生学习结果。

收稿日期: 2021-07-23

修回日期: 2021-11-23

基金项目: 国家自然科学基金项目(51774090); 黑龙江省省属本科高校基本科研业务费-东北石油大学创新基金项目(2019YDL-05); 黑龙江省教育科学“十三五”规划 2018 年度重点课题(GBB1318021)

作者简介: 郭玲玲(1978-), 女, 副教授, 研究方向为教育大数据、移动计算。

1 理论基础

1.1 在线学习行为

学习分为外显学习和内隐学习。在传统教学中,学习者在课堂环境中与学习者、教师和教学资源互动的行为称为学习行为。这样的学习行为只能通过教师的观察、视频存储和问卷调查来记录。但是这些方式不能每时每刻,十分全面地记录学生在课堂上的学习行为,会导致对学习行为分析造成一定的困难。始终缺少一手的实际数据,且没构成系统的理论分析体系是对于内隐的学习行为潜能的研讨的现状。由于在线课程的发展,在线学习应运而生。在线学习行为一般包括:在线学习时长、学生测试成绩以及学习者对教学视频、签到、教学安排等活动的点击行为。更多的是凭借学生自己的喜好自主进行选择,自然地呈现学习者的一些隐性学习行为^[5]。因此,对在线学习行为的研究变得越来越有意义,研究对象也逐渐从离线学习行为转向了在线学习行为。

1.2 机器学习

机器学习的目的在于通过算法让机器从海量历史数据中学习的规律,主动发现模式并用于推断未来的行为,是通过经历或数据来改良算法的研究^[6]。在学习分析中,机器学习的作用在于数据挖掘和数据解释,致力于完成传统教育无法完成的任务。机器学习包括:无监督学习、监督学习、半监督学习等^[7]。

1.2.1 无监督学习

无监督学习旨在挖掘整体数据集内部含有的隐含内容和规律,从而建立相应的模型实现对样本的剖析,主要针对的对象是没有标签的未标记数据集。

常见的无监督学习算法:聚类和降维。

(1)聚类:将数据分为多类的过程是根据数据的“相似性”。三种聚类算法比较如表 1 所示。

表 1 三种聚类算法比较

算法	伸缩性	簇形状	噪音敏感度	高维度	数据顺序影响	效果
层次聚类	差	任意形状	不敏感	好	敏感	较高
K-means 聚类	好	球形或凸型	敏感	好	不敏感	较高
DBSCAN	较好	任意形状	不敏感	一般	敏感	一般

经过比较,该文使用 K-means 聚类算法,优点在于算法较简易;收敛速度快;便于理解;当处理海量数据时有较高的可伸缩性;是很好的聚类算法即使数据集呈凸形和球形分布^[8]。

(2)降维:能够确保含有象征性的数据的特征或分布,同时将高维数据转化为低维数据。它通常可以用来可视化数据或中间过程,减少数据数量,提高其他

机器学习算法的效率。常用的降维方法有:主成分分析(PCA)、局部线性嵌入(LLE)、拉普拉斯特征映射^[8]。

由于 PCA 降维具有计算方法简单,便于实现;使用方差权衡信息量,不受数据集之外的要素影响等优势,因此,这里使用 PCA 降维方法。

1.2.2 监督学习

监督学习的训练集含有输入和输出,输入的是模型中的特征,输出的是目标。从已有的训练数据集中学习出来特定的函数(模型参数),当新数据到来时,能够依据该函数推测结果。监督学习常见的是分类问题,根据模型将一切输入映射为相应的输出,依据已有的训练样本得到一个最优模型,完成分类的目的根据的是对输出进行判断,模型应具备对随机输入进行分类的能力。

普遍的监督学习算法有统计分类和回归分析两种。统计分类的目标变量是标称型的,而回归分析的目标变量是连续数值型的。因此对于在线学习行为分析采用回归分析进行数据分析^[9]。回归分析常用的算法有线性回归和逻辑回归。线性回归具有建模速度快,可以根据系数给出每个变量的理解和解释;对异常值很敏感的优点,因此采用线性回归进行分析数据。

2 基于 K-means 聚类和线性回归建模

2.1 K-means 聚类

K-means 聚类算法先随机设定 k 个聚类中心,根据距离把样本点划分到所属类别,依次计算每个样本点到每个聚类中心的距离,并且选取最小距离,对每个样本点进行类别划分。之后在 k 个划分的类别中求出每个类别的质心位置,以该位置作为新的聚类中心,不停地迭代,直到聚类中心不再发生变化为止^[10]。K-means 聚类的具体流程如图 1 所示。

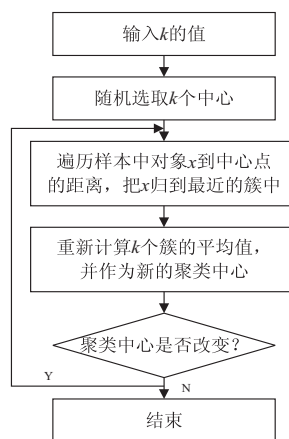


图 1 K-means 聚类具体流程示意图

2.2 PCA 降维

在 K-means 聚类分析中,会遇到特征维度比样本

数量多得多的状况,导致效果不一定好。一是由于冗余特征可能会导致一些噪音,影响计算的结果;二是因为无关的特征会造成计算量加大,消耗时间和资源。因而要对数据进行降维,缩小特征空间的维度,抽取子空间的数据来最优地表明实验所需的数据,从而减少参数预估的误差。

样本经过皮尔森系数筛选以及一些手动筛选剩余12个特征,特征维度较大,采取PCA降维对样本中的特征进行降维。具体运用机器学习sk-learn模块中的PCA函数来实现对特征进行降维。在聚类中删除了相关性弱的特征,仅用到6个特征;但在画图时,为了便于观察使用了全部特征来进行PCA降维,降至2维。

2.3 线性回归

建立模型之后,可以将问题转换为线性回归问题。由于线性回归算法简单,易于实现,这里选择线性回归算法分析学生不同的行为特征和成绩的关系。实验中对聚类结果的每个簇内特征的均值和方差进行分析后,发现簇内方差较小,均值差距明显;但由于数据规模较小,且较多数据方差较小,很难使用多变量非线性回归的方法处理,因此,选用单变量回归方法,分析单一特征对于成绩的关系,具体分析作业得分,观看视频时长,课程测试得分,访问数得分,任务点完成百分比,期末前综合成绩等对最终成绩的影响关系。

在分析单一特征对于成绩的关系时,对聚类后的6个点进行回归作图和折线作图。在回归作图中通过观察最小二乘回归模型的斜率以及阴影部分的面积(表示斜率的可能范围)可以知道该特征对成绩的影响趋势;在折线作图中,可以根据波动情况知道该特征对成绩影响的稳定性。

3 实验详细设计与实现

针对该校计科16级同学的《操作系统》课程数据,通过K-means聚类算法和线性回归算法,使用Python语言进行编写,实现学生在线学习行为数据分析^[11],主要功能与技术模块包括数据预处理、数据聚类分析、学习成绩与特征的线性回归分析^[12],如图2所示。

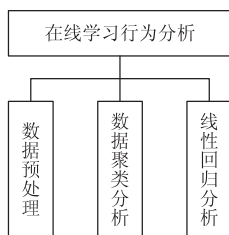


图2 主要功能模块示意图

3.1 数据预处理

(1)对采集的数据进行清洗。

将现有的学习数据通过删除、填充、重命名转换成算法可以分析的数据格式。

例如:学习者观看视频情况中存在学习者观看视频时长为空值,将这些空值以该列均值补全;有些学生因特殊情况退出本次课程学习,导致接下来学习得分全为空值,将这些数据进行删除。

(2)进行特征筛选。

①具有共线性的特征只选择其中一个,关键代码如下:

```
def delete_correlation(df, threshold=0.99):
    check_df = df.copy()
    if "学号账号" in df.columns:
        check_df = df.drop(["学号账号"], axis=1).copy()
    corr=np.abs(check_df.corr().values)>threshold
    to_delete=[]
    for i in range(corr.shape[0]):
        if np.sum(corr[i,i+1:])>1:
            to_delete.append(check_df.columns[i])
    cols=[col for col in df.columns if col not in to_delete]
    return df[cols]
```

②标准差小于0.5的删除,关键代码如下:

```
def std_analysis(df, threshold=0.5):
    std_filter = ['学号账号']
    process_col = [col for col in df.columns if col not in std_filter]
    for col in process_col:
        if df[col].std()>threshold:
            std_filter.append(col)
    return df[std_filter]
```

③计算皮尔森系数,系数高于0.5的特征保留,关键代码如下:

```
def corr_analyse(df, score_col, thres=0.5):
    strong_feature = []
    corr_dict = {}
    for col in df.columns:
        if col not in ['学号账号', 'UID', '表6-综合成绩', '表5-综合成绩']:
            corr = pearson(df, col, score_col)
            if corr>thres:
                strong_feature.append(col)
                corr_dict[col] = corr
    return strong_feature, corr_dict
```

经过皮尔森算法的筛选,学习行为特征包括:期末前综合测试成绩、五级制等级(ABCDE)、考试得分、课程视频得分、课程视频进度、课程测验得分、视频观看时间、任务完成百分比等^[13]。

这里以课程考试统计情况表为例说明一下统计情况,如表2所示。

表 2 课程考试统计情况

学号 账号	uid	...	期末前综合测试		
			成绩	提交时间	状态
160702140101	39927949	...	71.0	2018-12-13 20:18:16	完成
160702140103	39927951	...	86.3	2018-12-13 20:16:57	完成
160702140105	39927953	...	95.4	2018-12-13 20:34:33	完成

筛选过后的特征皮尔森系数如表 3 所示。

表 3 特征皮尔森系数统计

特征	皮尔森系数
期末前综合测试-成绩	0.5016168925685115
五级制等级 (ABCDE)	0.9671398346539911
考试得分	0.5016168925685115
课程视频得分	0.9750707687000391
课程视频进度	0.975058958888494
课程测验得分	0.5176992813526481
表 3-视频观看时长	0.5524206177231473
表 3-任务完成数	0.9772199902388473
表 3-任务点完成百分比	0.9772202163920058

3.2 实验结果

(1) 聚类分析。

聚类时,仅使用了每个子表的总计特征,其中相关性弱的特征被删除,用到的特征包括:作业得分、观看视频时长、课程测试得分、任务点完成百分比、期末前综合成绩等。画图时,使用全部特征进行 PCA 降维,降到 2 维,进行描点。

表 4 簇内特征均值统计

聚类中心	成绩	聚类个数	视频观看 时长	访问数	任务点完成 百分比	期末前综 合成绩	作业得分	课程测试
4	47.66	9	100.00	80.11	20.43	74.89	12.58	76.63
5	50.48	8	116.04	63.59	22.61	85.99	13.21	80.03
0	72.72	9	401.32	80.07	68.32	87.71	13.75	78.98
3	84.65	16	329.77	88.11	97.61	90.78	14.44	77.09
2	86.37	27	523.07	91.19	97.97	80.99	14.32	79.12
1	91.27	40	564.3	95.23	99.74	94.72	15.2	81.23

(2) 回归分析。

以“任务点完成百分比”特征为例进行说明。图 5 统计六个簇的簇内平均任务点完成百分比,上子图为回归分析图,下子图为折线图。

从图 5 中可以看出:

①上子图斜率明显,且阴影面积较窄,可得出任务

①PCA 降维结果如图 3 所示。

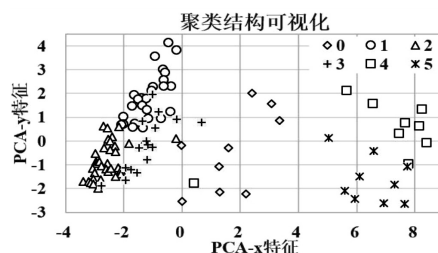


图 3 降维结果展示图

聚类时通过 PCA 降维方法将聚类使用的特征映射到 x 轴,y 轴为最后的得分,并且将不同的簇表示成不同的颜色。

②聚类分析结果如图 4 所示。

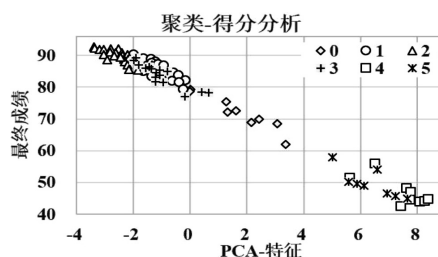


图 4 聚类结果可视化

通过画图可以看出使用筛选后的特征进行聚类之后,在成绩层面有很好的区分度^[14]。聚类结果的每个簇内,对统计特征的均值和方差进行分析可以得出,簇内方差较小,均值差距明显,聚类效果合理^[15-16]。因此统计每个簇内的特征均值,如表 4 所示。

结合统计表与可视化可以看出,对于视频观看时长、访问数、任务点完成百分比和期末前综合成绩四个特征,获得的经验越高,最终成绩越高;而对于作业得分和课程测试两个特征,结合 6 个聚类群体来看,对最终成绩影响不大。

完成百分比得分与最终成绩正相关线性关系明显;

②下子图波动较小,且有明显上升趋势,与上子图得出相同结论。

3.3 实验结果分析

通过对实验结果可视化进行的分析得出了各个特征对于最终成绩的影响程度,基于影响程度响学生以

及教师给出以下建议。

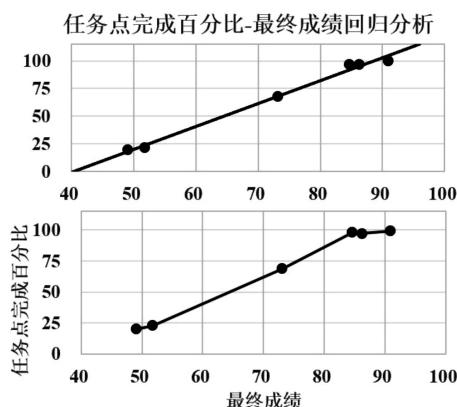


图5 任务完成百分比-最终成绩回归分析

(1)从学生的角度来看:

①从影响比较大的因素(观看视频时长、平台访问数、任务完成百分比、期末前综合测试)观察:其中影响最大的是任务完成百分比,其次是视频观看时长,然后是访问数,最后是期末前综合测试。应多花些时间去完成老师布置的任务;对于老师在平台上发布的视频应该及时观看,并且掌握其中的知识点;应多次进入在线教学平台去巩固自己的知识,预习或复习课程知识点;对于期末前综合测试应给予一定的重视积极复习。

②从影响比较小的因素(作业得分和课程测验得分)观察:只要认真对待课后作业和课程测验即可,不必在上边浪费太多时间。

(2)从教师的角度来看:

①从影响比较大的因素(观看视频时长、平台访问数、任务完成百分比、期末前综合测试)观察:首先发布任务之后应该督促学生及时完成,并且给予一定的指导;多发布一些与课程相关的视频,以便于学生观看学习。

②从影响比较小的因素(作业得分和课程测验得分)观察:应该尽量减少布置作业的数量,进行适当的课程测验即可。

4 结束语

通过使用机器学习算法中的K-means聚类算法来对在线课程教学数据进行分析。实验数据来源于东北石油大学16级计算机科学与技术学生使用在线平台学习《操作系统》产生的数据,对在线教学数据的分析取得了阶段性的成果。具体完成的研究工作如下:

(1)对学习进行聚类建模。

使用该模型能根据学生的学习行为对学习进行聚类。教师可根据聚类结果给予学生更为优质更为科学的学习资源,对不同类型的学生给予不同的指导建议。

(2)对学生学习行为进行分析。

在线平台给予学生的在线活动很多,导致衡量学生学习成效的考虑因素很多,对学生最终学习成绩影响因素进行筛选。从在线平台上学生学习行为中选取对学生最终成绩产生较大的行为,并确定筛选出来的影响因素的重要程度。给予教师指导意见,以便于教学帮助学生提高学习能力。

参考文献:

- [1] 李乡儒,梁惠雯,冯隽怡,等. 在线教育平台中个性化学习资源推荐系统设计[J]. 计算机技术与发展,2021,31(2): 143-149.
- [2] 李建伟,苏占玖,黄赞茹. 基于大数据学习分析的在线学习风险预测研究[J]. 现代教育技术,2018,28(8):78-84.
- [3] ZHENG Jianxing, WANG Suge, LI Deyu, et al. Personalized recommendation based on hierarchical interest[J]. Information Sciences, 2019, 479: 55-75.
- [4] 周海波. 基于自适应学习平台促进学生个性化学习的研究[J]. 电化教育研究, 2018, 39(4): 122-128.
- [5] 彭丽华,王陈欣,刘雪宁,等. 大数据视角下的在线学习行为研究[J]. 中国教育信息化, 2021(7): 1-5.
- [6] 李想. 基于机器学习的在线学习行为投入评测模型研究与实现[D]. 上海:上海师范大学, 2021.
- [7] ETHEM A. Machine learning: the new AI[M]. [s. l.]: MIT Press, 2016: 1-30.
- [8] 黄吉. 一种K-means聚类改进算法研究及应用[D]. 武汉:湖北工业大学, 2018.
- [9] STRANG K. How student behavior and reflective learning impact grades in online business courses[J]. Journal of Applied Research in Higher Education, 2016, 8(3): 390-410.
- [10] 李涛. 大数据时代的数据挖掘[M]. 北京:人民邮电出版社, 2019: 13-47.
- [11] 张学刚,李倩雯,王笑丹,等. 智慧学习环境下学生学习行为对比分析研究[J]. 现代职业教育, 2021(30): 32-33.
- [12] 孙众,吕恺悦,骆力明,等. 基于人工智能的课堂教学分析[J]. 中国电化教育, 2020(10): 15-23.
- [13] 李福顺. 教育数据挖掘的在线学习机制研究[J]. 微型电脑应用, 2021, 37(6): 163-165.
- [14] ZHANG S, LIU Q T, CHEN W L, et al. Interactive networks and social knowledge construction behavioral patterns in primary school teachers' online collaborative learning activities[J]. Computers & Education, 2017, 104: 1-17.
- [15] 郭永坤,章新友,刘莉萍,等. 优化初始聚类中心的K-means聚类算法[J]. 计算机工程与应用, 2020, 56(15): 172-178.
- [16] 蔺小清. 大数据时代K-means聚类算法应用于在线学习行为研究[J]. 电子设计工程, 2021, 29(18): 181-184.