基于大数据挖掘和用户画像的在线课程学习效果评价模型研究

邱 斌
宁波职业技术学院 浙江 宁波 315800

摘 要：本文主要介绍了在大数据的背景下教育领域特别是在高校中对在线课程平台中积累的海量的过程性的学习行为数据进行分析，使用数据挖掘算法和用户画像技术对学习效果进行更为客观有效的评估，对学生建立标签结构体系，并进行学生画像，完成学生在线效果评价模型的构建，实现在线学习效果个性化评价，为学生和教育者提供了改进教学效果的意见，从而提在线学习平台的课程的学习效果。

关键词：数据挖掘、用户画像、学习效果评价。

引言

由于在线学习平台学习效果分析研究方面，目前在理论和应用方面都已经取得一系列成果。国外关于网络学习行为分析的研究主要是应用性的研究，且主题更加新颖、研究更加深入。例如，Pavlo D.Antonenko等通过层次聚类方法和非层次聚类方法分析来自在线学习环境服务器日志点击流数据，识别出学习成绩优秀学生的学习行为特征[1]。Alil等人使用他们自研的LOCO-Analyst系统，该程序以图文混合的方式显示学生的学习反馈情况，生成个性化的学生评价数据提供给任课教师参考[2]。

1 2014 年以来，越来越多的国内研究者开始关注数据挖掘技术在教育领域的应用研究。

这类的研究主要集中在对学习者的学习行为的分析和学习效果的预测、教辅软件系统的开发、上机考试结果的数据分析挖掘等方向上。张晓军等人使用聚类分析应用到学生成绩评价，对特定的某个专业的学生期末考试成绩进行了分类与评估[4]。樊同科利用因子分析技术为学生奖学金和就业岗位推荐等方面进行学生综合评价方案的优化[5]。韩宇等人通过对高校的课程进行因子分析，在传统的评价方案的基础上提出了一种学生学习能力评价的新方法[6]。齐宗会在学情分析中应用了K-Means聚类算法对学生进行多个维度的评价，排除了考试难度等多个影响学生评级效果的因素的干扰[7]。另外，国内越来越多的学者们对个性化教学评价进行研究并取得了相关成果。陈雄辉等人通过对多种教学评价参数进行量化分析，构建一套科学实用的评价个性化学习课堂教学的指标体系[8]。吴守蓉等人的对MOOC课程从学习者的视角，从选课动机、学习行为、满意度等方面，通过问卷调查和数据统计分析出教学改进策基于大数据的数据挖掘分析[9]。

虽然在国内外都对通过大数据挖掘的教学效果评价都进行了很多的研究和应用实践。但是，很多研究者通过数据挖掘的各种算法并且从设计的多维的评价指标和方法来衡量学生的学习效果，并进行个性化学习效果评价中的应用却非常少。

近年来，用户画像技术在很多领域例如在线精准营销、信息和商品的个性化推荐等领域中得到的广泛的应用。但是在高等教育领域，对于大学生的数据采集、处理以及用户画像构建等方面还处于起步阶段。用户画像技术在个性化学习效果评价领域的应用还处于起步阶段。使用用户画像技术对学生进行多个维度的特性描述，然后使用描述的结果对学习效果进行多维度的评估，即个性化的学习效果评价。

在教育领域中，针对不同的学习者建立用户画像以达到精准教学的目的是目前教育大数据的重要研究内容。本文研究在线学习平台的学习效果评价，是在大数据分析的基础上对学生进行用户画像，将学生的个性化数据加入加入到综合评价中去，改变了以往机械的通过简单的成绩分数等单个维度的评价标准，提供了学生学习和教师教学效果的全面的评价体系。

改变了单纯的成绩导向的评价系统，能够更好促进学生的全面发展。

2 在线学习评价方法建模

2.1 数据采集

本文以我校计算机应用技术专业的专业核心课程作为研究对象。从在线学习平台中每天产生的数以万计的行为数据中结合数据统计和数据处理软件，分析出学生在在线学习平台上进行学习行为数特征。

本文研究的数据集合由学生的线上的基本信息数据和在线学习的行为数据两个部分组成。时间跨度为2020年1月至2021年12月之间的累计4个学期的所有专业课程中的所有选课学生的在线学习的行为数据。通过在线学习平台提供的学生行为日志导出功能，将学习行为数据导出，并将这些数据保存在用于分析的mysql数据库中。在线学习平台上产生的行为数据记录，其中每条记录代表着每一个事件的记录。一个事件是指学生登录在线学习平台后进行操作痕迹信息，包括事件ID、事件发生时间、用户ID、课程资源ID、课程资源内容、行为记录的内容描述、事件类型、访问系统的方式、用户IP地址等信息。

2.2 数据预处理与分析

数据预处理的主要目的对获取的数据集中的不完整的部分和异常的数据进行处理。在进行后续的数据挖掘之前，对采集到的学生基本信息和行为相关数据进行预处理是非常重要的一个步骤。本文采用Spark大数据分析技术框架并结合scala程序对库中的数据进行初步分析，找到其中的异常数据和缺失值记录。

在教学过程由于实际情况出现的异常数据可以考虑保留。对于缺失值情况可以分为两种，如果该缺失值可以从线下的纸质文档或者其他教学业务系统中查询得到，可以考虑将缺失的数据填补后继续使用，如果通过查询得到则将该条数据记录删除。如果某个或某类事件的缺失数据较多，则可以使用均值用K近邻插补的方式进行填充。

在获取原始的在线学习平台事件数据的基础上可以，进行初步的数据分析统计。

2.2.1 单次在线学习时间的计算。虽然每次登录的时间可以非常容易从登录事件的时间记录中读取，但是在线学习平台并没有记录学习者的退出学习平台的时间。所以，可以按照时间排序事件记录，从学习者登录在线学习平台后统计在一段时间内该学习者最后的活动为止，两个时间的差值就是认为是该次学习的时长。也可以将该段时间内，学习者观看的视频的时长作为统计学习时长一个重要依据。

2.2.2 学习内容分析。通过分析事件日志，可以分析和统计单次学习时长内学习者的学习的内容包括课件、视频、文档等电子资料。

在线学习平台的行为日志记录一般是高维的数据，另外记录条数也十分庞大，数据之间的关联也异常复杂。高维数据处理难度大，维度太低也也会影响用户画像的准确性。为了方便进行后期的数据处理分析，减少无效数据的干扰，就需要对原始的平台事件日志数据进行预先的分析处理，只保留必要的字段，满足用户画像数据即可。由于在线学习平台记录了学习者的个人隐私等信息，所有涉及个人的敏感信息就在处理之前就必须要进行脱敏处理，可以使用别名以及使用密钥加密后的HASH算法处理敏感数据。

2.3 数据挖掘及过滤

数据挖掘和过滤是用户画像的核心过程。用户画像是对学生个人以及学习特征的总体概要描述。本文从学生学习行为、学生的基本信息数据、学生标签集这三类数据来生成用户画像的标签，并且以此来创建该用户的画像，从而实现个性化的学生学习和教师教学评价任务。数据挖掘的方法有很多种，常用的方法有聚类、分类、关联规则、决策树等。

本文使用K-means为基础结合层次聚类的聚类集成算法来构建学生分类模型。从学生的多维度属性生成学生的个性化学习特征。最后利用投票法对上述的分类模型的分类结果进行集成，得到某个学生最终的类别。

2.4 标签的提取与重组

标签的提取与重组是用户画像的最后一个步骤。这个环节处理效果会对用户画像的准确性产生最直接的影响，如对某个标签的权重的修改就会对用户画像模型产生影响。

在数据预处理与分析后得到的数据有三类。

第一类是学生的基本信息包括了用户id、学生班级、年龄、性别、所学专业等信息。 第二类是学生在线学习总时长，在某一门课程的登录次数、课程资源下载和微课资源观看的数量，这类数据用于后期分析挖掘学生的学习投入度。

第三类是课程资源和附件资源的类型。通过对资源类型的数据挖掘可以得到学生学习内容偏好的信息。

通过上述的三类数据，可以构建结构化的标签体系。学生的用户画像的标签集由三个维度组成，分别是学生基本特征、学习投入度以及媒体偏好。

2.5 用户画像

学习者的用户画像的标签有个人基本特征、学习状态和学习内容偏好三个维度表示。

个人基本特征是学习者的唯一标识，通过对数据记录的多个维度使用数据挖掘算法得到。

学习状态包括了学习频度和行为投入两个指标。

学习行为投入分为高、中、低三种投入类型。行为投入数据描述由课程资源的观看下载量、观看课程视频的数量以及登录次数一起构成。

学习频度包括了每门课程的学习时长、单位时间内

学习时长以及学习持久性。学习持久性是学生在线学习的持续时间长度，可以分为高，中，低三种持久性，由单次登录系统后学习持续的时间长度来划分。

学习内容偏好是学生在在线学习课程更偏向于学习哪种类型的资源或媒体。从事件日志中可以挖掘内容偏好信息包括文本(包括课件、网址等)，图片和视频三类资源媒体。特征指标是三种媒体资源的学习的总次数。

由上述各维度的用户画像建模，可以得到学生在线学习效果的评价。上述的在线学习多维度的特征指标用聚类算法处理后生成对学生在线学习效果评价结果，分为{A，B，C，D，E}五个等级。

2.6 用户画像模型准确度验证

为了验证上述模型生成的学生个性化学习效果评价的可信度，本文对本校教务部门提供2020和2021两个学年合计四个学期所有课程的成绩与用户画像模型分析得到的学生学习效果评价结果进行对照，得到模型的准确度。根据对照结果，分析出现模型分析偏差的原因，并对数据挖掘算法进行改进，进一步提高模型分析的准确度。

3 总结

本文研究的基于大数据挖掘用户画像学习评价模型能够通过爬取在线学习平台上的学生基本数据和学习过程性数据对学生学习效果进行有效的预测与评价，为课程教学考核与评价以及学校的教学管理提供了有力的数据支持。另外，基于学生用户画像的个性化的学习效果与评价机制完成学生个性化描述任务。根据用户画像生成在线学习的个性化学习效果评价报表，学生能够通过可视化评价报表了解自己的学习的薄弱环节和今后的改进的方向。

任课教师也可以通过评价结果报表对学生的学习情况有一个多角度掌握，改进教学方法，调整教学内容，提升教学效果。

参考文献

[1]Aldowah H, Al-Samarraie H, Fauzy W M. Educational data mining and learning analytics for 21st century higher education: A review and synthesis[J]. Telematics and Informatics, 2019, 37: 13-49.

[2]Ali L, Hatala M, Gašević D, et al. A qualitative evaluation of evolution of a learning analytics tool[J]. Computers & Education, 2012, 58(1): 470-489.

[3]Gonçalves A L, Carlos L M, da Silva J B, et al. Personalized Student Assessment based on Learning Analytics and Recommender Systems[C]//2018 3rd International Conference of the Portuguese Society for Engineering Education (CISPEE). IEEE, 2018: 1-7.

[4]张晓军,李珊珊,杨树生.基于MATLAB的因子分析与聚类分析在学生成绩评价中的应用[J].聊城大学学报(自然科学版),2015,28(02):71-75.

[5]樊同科.基于因子分析法的计算机专业学生成绩综合评价方法研究[J].软件导刊(教育技术),2018,17(01):11-13.DOI:10.16735/j.cnki.jet.2018.01.007.

[6]韩宇,包红.因子分析法在学生成绩综合评价中的应用[J].数理医药学杂志,2016,29(05):785-786.

[7]齐宗会.主成分聚类分析法在综合评价学生成绩中的应用[J].太原城市职业技术学院学报,2016(08):182-183.DOI:10.16227/j.cnki.tycs.2016.0720.

[8]陈雄辉,刘晓,赵丹丹,刘繁华,刘博,于淑霞,崔慧洁.教育信息化2.0时代个性化学习课堂教学评价指标体系的构建[J].广东技术师范大学学报,2020,41(05):28-33+41.DOI:10.13408/j.cnki.gjsxb.2020.05.005.

[9]吴守蓉,崔璨,汪琼.基于学习者视角的MOOC教学评价与改进——以北京大学“教你如何做MOOC”课程为例[J].中国大学教学,2016(10):68-76.