文章编号:1003-6180(2020)03-0057-05

教育数据挖掘和学习分析研究进展

杨高明,方贤进,葛 斌,张 玉
(安徽理工大学 计算机科学与工程学院,安徽 淮南 232001)

摘 要:综述教育数据挖掘和学习分析在高等教育中的应用,探讨计算机支持的学习分析、计算机支持的预测分析、计算机支持的行为分析、计算机支持的可视化分析的各种学习问题以及所使用的数据挖掘技术,提出应用教育数据挖掘和学习分析帮助高等院校做出更好决策的方案.

关键词:教育数据挖掘;机器学习;高等教育;学习分析
DOI:10.13815/j.cnki.jmtc(ns).2020.03.014
[中图分类号]TP309 [文献标志码]A

A Survey of Educational Data Mining and Learning Analysis

YANG Gaoming, FANG Xianjin, GE Bin, ZHANG Yu
(School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China)

Abstract: This article reviews the application of educational data mining and learning analysis in higher education. At the same time, we also discuss the various learning problems of computer-supported learning analysis, computer-supported predictive analysis, computer-supported behavior analysis, computer-supported visual analysis, and data mining techniques used. Our purpose is to apply educational data mining and learning analysis to help colleges and universities make better decisions.

Key words: educational data mining; machine learning; higher education; learning analysis

教育数据挖掘(Educational data mining, EDM)和学习分析(learning analytics, LA)可以连续收集、处理、报告和处理数字数据,改善教育过程,重塑现有的教学模式,为教师和学生的交互提供新的解决方案和更多个性化、适应性和互动性的教育环境,提高学习成果,优化机构管理水平,对教师和学生做综合评价. 教育数据挖掘主要解决两个问题:人们如何使用EDM和LA解决教育方面的实际挑战,哪种数据挖掘技术最适合这些问题. 本文从技术角度阐述EDM和LA在高等教育中的最新进展,目的是为机器学习领域从事基于数据挖掘的同行提供参考.

收稿日期:2020-01-15
基金项目:国家自然科学基金项目(61572034);安徽省高校自然科学基金项目(KJ2019A0109);安徽省重大科技专项基金项目(18030901025);教育部产学研项目(201901051005);安徽理工大学研究生核心课程(2019HX006);安徽省信息安全卓越工程师培养计划(2019)
作者简介:杨高明(1974-),男,安徽临泉人. 副教授,博士,主要从事机器学习和学习分析研究;方贤进(1970-),男,安徽舒城人. 教授,博士,主要从事信息安全和数据挖掘研究;葛斌(1973-),男,安徽潜山人. 教授,博士,主要从事信息安全和数据挖掘研究.

1　教育数据挖掘分类

EDM和LA分为四个方面:计算机支持的学习分析(computer-supported learning analytics,CSLA)、计算机支持的预测分析(computer-supported predictive analytics,CSPA)、计算机支持的行为分析(computer-supported behavioral analytics,CSBA)和计算机支持的可视化分析(computer-supported visualization analytics,CSVA).CSLA的研究主要集中在使用数据统计分析的方法,在课程环境中分析学生的信息搜索和协作学习行为.CSPA的研究集中在使用预测函数或连续变量改善学生的学习和表现以及评估学习材料的适当性.CSBA的研究主要关注发现学生行为和知识模型.CSVA的研究集中在可视化探索数据(使用交互式图表)方法上,从而突出显示有用的信息并产生准确的数据决策.

协同学习　EDM和LA通常用于处理与教学策略有关的问题,这些策略可以支持和增强合作学生之间协作过程,是衡量协作有效性的主要指标,其中学习平台中用户活动日志被用作推断学习者活动的主要工具,以适应特定人群行为和偏好.[1] C. VieiraP[2]研究了EDM对计算机支持协作学习在会议期间学生参与的影响,对学生参与元素做了可视化,并帮助他们在CSCL中更好地进行协作.Cerezo[3]使用慕课数据研究学生与LMS互动模式,以帮助教师更好地了解学生的各种学习特征,从而帮助他们识别有学习困难的学生.

社会网络分析　使用EDM和LA可以根据个人的学习活动以及他们在文化和社交环境中共同建立的关系进行社交网络分析,包括发现学术合作、评估社交关系、推荐学习课程等.Duval[4]通过收集用户行为的数据提供有关学习资源和活动的建议.

自学行为　EDM和LA通过调查学生对学习资源和自我评估练习的使用情况及其对他们的表现影响,为在线自我学习环境提供了一个有效解决方案.[5] R. L. Rodrigues[6]根据学生解决问题的进度来检查学习系统中的自我调节学习行为.A. Littlejohn[7]运用EDM来预测学习者如何根据目标设定和监控活动的证据将其智力能力转化为学术技能.

1.1　计算机支持的预测分析(CSPA)

数据挖掘通过评估学习材料、学生之间参与的相互作用来增强当前的教学经验,降低学生的辍学率和留级率.[8] S.　RizviB[9]认为,在学习环境中使用数据挖掘技术可以发现大量数据中隐藏的知识和模式,并预测结果或行为.A. A. Saa[10]认为,可以使用EDM和LA发现知识,帮助教师识别早期辍学的学生,并确定需要特别关注的人.

学习资料评估　数据挖掘为分析和研究学习管理系统数据提供了足够的方式,以提高高等教育的质量.T. Devasia[11]认为,数据挖掘可用于研究影响学生表现的主要属性,给学习者不同的教学支持.支持性反馈可以帮助教师了解学生对授课的反应,从而评估课程的有效性,帮助课程设计者、教师和机构进行决策.

评估和监督学生的学习　学生的学习评估和监控实践是高等教育的重要方面.学习效率考核包括评估和评估过程,这些过程可以帮助学生、教师、管理人员和高等教育机构的决策者做出决策.当代教育可以使用各种数据挖掘技术监测学生的表现,提供各种调查分析方法,发现教育系统中隐藏的信息,以便生成评分.W. Yaacob[12-13]认为,数据挖掘可用于识别学生行为和他们学习的方式,发现不良行为并进行学业分析,预测学生的成绩.S. Bharara[14]使用EDM和LA分析学生的学习行为,并警告处于危险状态的学生以免他们中断学业.Salas[15]通过分析学生的行为创建聚类,支持科学技能的学习,以增

强学生的科学能力，并利用从学生互动中收集的信息为学生提供必要的帮助，以提高他们的元认知能力.

辍学和留级　慕课为学生学习提供了方便，但退出在线课程的学生人数一直在增长，研究人员研究了各种退课的因素，发现这些因素在各个教育水平上都阻碍了学生的表现. Pradeep[16] 使用 EDM 研究分析了影响学生学习成绩的因素，预测学生辍学情况，识别出表现较差的学生. Cambruzzi[17] 发现学生的辍学状态具有很高的可预测性，平均准确性为87%，使用他们的预测结果以后，辍学率平均降低了11%.

1.2　计算机支持的行为分析(CSBA)

了解学生的学习行为模式可以更好地服务教师和学生，而利用数据挖掘技术可以很容易发现学生的行为模式. J. W. You[18] 发现，在远距离合作时，使用 EDM 和 LA 可以改善学生的学习体验. R F Kizilcec[19] 设计了一个学生模型，通过结合有关学生的知识、动机、元认知和态度等信息预测学习过程. C Angeli[20] 认为，EDM可用于评估学生在线活动与其最终成绩之间的关系来检测学生在网络环境中的不正常行为和活动.

1.3　计算机支持的可视化分析(CSVA)

教育数据可视化可以表示学生对学习任务的参与，帮助教师更好地了解学生的在线行为，简化复杂数据，跟踪在线教育系统中学生的交互信息.[21] CSVA将信息可视化技术与数据挖掘和知识表示相结合，提供有关活动的个人行为可视化分析，以便研究者直观的观察研究结果.[22] O W Adejo[23] 认为，在高等教育评估系统中使用视觉数据挖掘可以使评估方法更灵活、更多样、更直观，从而提高学习效率. X Du[24] 研究了使用 EDM 从大数据集中提取有意义的知识和信息，并使用此信息发现对高级决策有用的隐藏模式和关系.

2　数据挖掘技术

2.1　分类

分类技术可以有效地为学生提供早期干预，特别是激发在特定活动或课堂上表现不佳的学生，并准确衡量该形式的效率. 分类是教育数据挖掘中常用的技术，它属于监督学习，既给定训练数据，对测试数据进行预测. 训练数据由输入输出对组成，训练数据通常表示为：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}.$$

分类模型首先根据训练数据训练分类器模型参数，然后根据测试数据对分类器模型参数进行微调，最终成为学习过程(学习系统). 分类模型构造好之后，新的预测数据就可以使用该模型进行预测.

分类模型已经用于预测学生的表现、学习成就、知识水平等，也应用于预测/防止学生辍学、检测在线课程/学习中有问题学生的行为上. B K Francis[25] 指出，分类技术可以通过准确预测学生在特定课程中的最终成绩，提高高等教育系统的质量. 分类技术的目标是检查参与水平，防止学生退出远程学习和在线学习课程[26]；评估学生对学习活动的参与度；持续评估学生的学习表现[27]；识别学习积极性不高的学生；确定学生是否会完成作业[28]；评估学生与学习材料的互动[29]. 此外，分类还用于提高学习的效率和有效性，为高等教育系统提供一些指导，从而改善整个决策过程. 可以说分类为决策者提供了更大的灵活性，使他们能够评估一组学生的表现和行为，从而确定该组中的每个成员学习中如何表现良好，即使他们的特定知识或能力不适合该任务，也可以执行该任务.

2.2　聚类和回归分析

高等教育聚类是学生分组的有效技术，它可以用来探索协作学习模式并提高保留率，使机构

能够在早期识别出具有学习风险的学生. 聚类属于无监督学习,目的是通过得到的类或者簇发现数据的特点或对数据进行处理. 聚类的核心是计算数据之间的距离. 在教育数据挖掘领域使用聚类主要是为了支持学生在不同学习情况下的互动[30],向相似的用户推荐活动和资源,根据访问页面的内容和学习特征找到具有相似学习特征的学生群体,帮助教育决策者尽早发现潜在的辍学者,并解决将新学生分配到他们不感兴趣的课程的问题. 此外,聚类可以使教育者从LMS日志中预测学生的学习结果,识别不良的学生行为,并通过监视学生之间的集体互动,支持教师评估学生建模合作过程的学习状况. 该技术还被用于支持学生掌握各种科学技能[31],发现常见的学习途径,了解学生之间的协作过程.

回归可以有效地用于预测,EDM研究人员经常运用几种回归技术来预测学生的学习成绩,并确定可以预测大学课程成败的变量.

3 结果与讨论

本文综述了EDM和LA在高等教育中的应用,探讨CSLA,CSPA,CABA和CSVA的各种学习问题以及所使用的数据挖掘技术. EDM/LA的应用可以带来巨大的好处,可以帮助高等教育机构开发更多以学生为中心的课程,实时预测学生的学习状态,推荐合适的课程.

数据挖掘技术可以使高等院校做出更好的决策,在指导学生更准确地预测未来和个人行为时提供更高级的计划,并使高等院校更有效地分配资源和人员. 使用EDM和LA可以改善学生的学习体会、学习成果、发现模式,预测学生的行为和成就.

CSPA通常使用不同的数据挖掘技术评估在线学习材料,根据学生的最终成绩来监控学生的表现. 分类是解决CSPA问题最常用的技术,其次是聚类. 分类和预测都被用来形成用于促进某些学习任务的学习模型;聚类技术可以基于学生的互动和学习困难的模式对学生进行分组来识别类似班级的对象,发现常见学习途径及不良的学生行为;可视化和统计技术可以提供学生学习的总体视图,突出有用的信息并支持整个决策过程.

CSBA主要关注使用 EDM 和 LA 使大学能够发现大型数据库中的隐藏模式,并以高准确度构建模型,为设计在线课程提供有效的解决方案. 聚类是解决与CSBA相关学习问题的最常用技术,它可以有效地识别与学生学习风格有关的隐藏模式,并发现不良的学生行为. 分类技术是第二常用的技术,主要用于构建和发展学生表现的预测模型. 再次是关联规则挖掘和可视化. 相关挖掘、因果数据挖掘和离群值检测使用最少.

CSVA使用不同类型概念图表示已知/未知概念,表示学生的知识水平以及帮助解决数据表示问题. 视觉数据挖掘技术使用较广泛,用于发现先前未知和隐藏的信息以及数据中的模式. 使用可视化技术可以提供数据的全面视图,以图形方式呈现LMS/CMS收集的复杂学生跟踪数据,识别有趣的子集. 这些结果可以揭示有价值的信息以及隐藏的见解、关联或关系,用来促进对学生在不同学习环境中的互动更深入了解,使决策者和系统开发人员能够有效地重新设计学习机会和课程.

参考文献

[1]Kurilovas E. Advanced machine learning approaches to personalise learning:learning analytics and decision making[J]. Behaviour & Information Technology,2019,38(4):410-421.

[2]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.

[3]Cerezo R,Sánchez-Santillán M,Paule-Ruiz M P,et al. Students' LMS interaction patterns and their relationship with achievement:A

case study in higher education[J]. Computers & Education，2016，96：42-54.
[4]Slater S，Joksimović S，Kovanovic V，et al. Tools for educational data mining：A review[J]. Journal of Educational and Behavioral Statistics，2017，42(1)：85-106.
[5]杨文君. 数据挖掘在教学中的应用分析[J]. 牡丹江师范学院学报：自然科学版，2005(3)：29-30.
[6]Rodrigues R，Ramos J，Silva J，et al. Forecasting students' performance through self-regulated learning behavioral analysis[J]. International Journal of Distance Education Technologies，2019，17(3)：52-74.
[7]Littlejohn A，Hood N，Milligan C，et al. Learning in MOOCs：Motivations and self-regulated learning in MOOCs[J]. The Internet and Higher Education，2016，29：40-48.
[8]杨晓华. 数学分析合作性学习的要素和作用[J]. 牡丹江师范学院学报：自然科学版，2009(3)：64-65.
[9]Rizvi S，Rienties B，Khoja S. The role of demographics in online learning：A decision tree based approach[J]. Computers & Education，2019，137：32-47.
[10]Saa A. Educational data mining & students' performance prediction[J]. International Journal of Advanced Computer Science and Applications，2016，7(5)：212-220.
[11]Devasia T，Vinushree T，Hegde V. Prediction of students performance using Educational Data Mining[C]. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)，2016：91-95.
[12]Dutt A，Ismail M，Herawan T. A systematic review on educational data mining[J]. IEEE Access，2017(5)：15991-16005.
[13]Yaacob W，Nasir S，Yaacob W，et al. Supervised data mining approach for predicting student performance[J]. Indonesian Journal of Electrical Engineering and Computer Science，2019，16(3)：1584-1592.
[14]Bharara S，Sabitha S，Bansal A. Application of learning analytics using clustering data Mining for Students' disposition analysis[J]. Education and Information Technologies，2018，23(2)：957-984.
[15]Salas D，Baldiris S，Fabregat R，et al. Supporting the acquisition of scientific skills by the use of learning analytics[C]. International Conference on Web-Based Learning，2016：281-293.
[16]Pradeep A，Das S，Kizhekkethottam J. Students dropout factor prediction using EDM techniques[C]. 2015 International Conference on Soft-Computing and Networks Security (ICSNS)，2015：1-7.
[17]Cambruzzi W，Rigo S，Barbosa J. Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach[J]. Journal of Universal Computer Science，2015，21(1)：23-47.
[18]You J. Identifying significant indicators using LMS data to predict course achievement in online learning[J]. The Internet and Higher Education，2016，29：23-30.
[19]Kizilcec R，Pérez-Sanagustín M，Maldonado J J. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses[J]. Computers & Education，2017，104：18-33.
[20]Angeli C，Howard S，Ma J，et al. Data mining in educational technology classroom research：Can it make a contribution？[J]. Computers & Education，2017，113：226-242.
[21]Noroozi O，Alikhani I，Järvelä S，et al. Multimodal data to design visual learning analytics for understanding regulation of learning[J]. Computers in Human Behavior，2019，100：298-304.
[22]Rodrigues M，Isotani S，Zárate L E. Educational data mining：A review of evaluation process in the e-learning[J]. Telematics and Informatics，2018，35(6)：1701-1717.
[23]Adejo O，Connolly T. Predicting student academic performance using multi-model heterogeneous ensemble approach[J]. Journal of Applied Research in Higher Education，2018，10(1)：61-75.
[24]Du X，Yang J，Shelton B，et al. A systematic meta-review and analysis of learning analytics research[J]. Behaviour & Information Technology，2019(9)：1-14.
[25]Francis B，Babu S. Predicting academic performance of students using a hybrid data mining approach[J]. Journal of Medical Systems，2019，43(6)：162.
[26]Burgos C，Campanario M，Pena D，et al. Data mining for modeling students' performance：A tutoring action plan to prevent academic dropout[J]. Computers & Electrical Engineering，2018，66：541-556.
[27]Rawat K，Malhan I. A hybrid classification method based on machine learning classifiers to predict performance in educational data mining[C]. Proceedings of 2nd International Conference on Communication，Computing and Networking，2019：677-684.
[28]Dragulescu B，Bucos M，Vasiu R. Predicting assignment submissions in a multi-class classification problem[J]. TEM Journal，2015，4(3)：244.
[29]Paiva R，Bittencourt I I，Tenório T，et al. What do students do on-line？Modeling students' interactions to improve their learning experience[J]. Computers in Human Behavior，2016，64：769-781.
[30]Ramanathan L，Parthasarathy G，Vijayakumar K，et al. Cluster-based distributed architecture for prediction of student's performance in higher education[J]. Cluster Computing，2019，22(1)：1329-1344.
[31]Salas D J，Baldiris S，Fabregat R，et al. Supporting the acquisition of scientific skills by the use of learning analytics[C]. International Conference on Web-Based Learning，2016：281-293.

编辑：琳莉