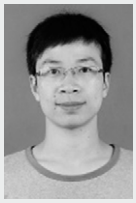
大数据引领教育未来：从成绩预测谈起

Big Data Drives a New Epoch of Education: A Case Study of Academic Performance Prediction

吕红胤，女，电子科技大学副研究员，主要研究方向为教育大数据理论与实践研究、社会圈层研究。



连德富，男，电子科技大学讲师、教育大数据研究所副所长，主要研究方向为机器学习、时空数据挖掘、推荐系统、教育数据挖掘。在ACM Trans.、KDD、Ubicomp、ICDM、WWW等国际顶级期刊和会议上发表论文10余篇。

聂敏，男，电子科技大学教育大数据研究所博士生，主要研究方向为大规模分布式计算、教育数据挖掘。现任成都寻道科技有限公司总经理，致力于教育大数据平台级产品研发，有多年大数据相关技术经验。

夏虎，男，电子科技大学副研究员、教育大数据研究所所长，主要研究方向为数据挖掘、社会网络、文本挖掘。

周涛，男，电子科技大学教授，主要研究方向为统计物理与复杂性科学，发表SCI论文200余篇，引用12 000余次，H指数为53。

doi: 10.11959/j.issn.2096-0271.2015045

近年来，大数据已经在教育领域的管理与引导等诸多方面被广泛运用。例如，智能教学系统（ITS）[1]基于与学生间交互的日志数据进行个性化知识诊断，分析学生的知识掌握情况，发现学生的薄弱点，从而自适应地帮扶学生更好地获取知识和技能[2,3]。卡耐基公司（Carnegie Learning）的“认知导引”系统便是一个典型的ITS[4]，它根据学生对先前问题的回答情况制定后续的提问内容。这样，就可以找出学生的问题并深入了解它们。经过严格测试，发现使用该系统的学生要比接受传统教学的学生节约12%的学习时间。国外的edX、Coursera、Udacity和国内的学堂在线等多家大规模在线课堂平台，围绕在线教育中高辍学率的严峻问题，基于学生人口统计学数据和学生注册课程、观看视频、完成课后作业、参与论坛讨论等行为数据，旨在发现影响学生辍学的重要因素，从而制定相应的干预策略引导学生，降低在线教育的辍学率[5]。保罗·艾伦实验室致力于自动化答题的研究：从题目中抽取知识和前提条件，基于在训练集上构建的知识图谱，利用多种统计推断和逻辑推理的方法来选择或者生成可能准确的答案[8,9]。针对该任务，该实验室在2015年10月发起了一项名为“你的模型比8年级学生更聪明吗”的大数据竞赛。更重要的是，包括中、美、日在内的国家均已设立了答题机器人的国家级重大项目，计划让机器人在不久的将来参加高考，在3~5年内考上“一本”。而且，目前该项目已经取得可喜进展。

除此以外，传统中小学教育和高等教育中积累的人口统计学信息、过往考试成绩、缺旷课、问卷调查等数据也曾被用于分析与学生综合绩点、能否顺利毕业等因变量之间的关系，并且构建相关的预测模型[10]。基于预测模型，教育管理者便可以优先找出未来可能需要重点关注的学生。然而，这些数据要么可能只是来源于小部分学生的问卷调查，要么数据的字段数太少。更重要的是这些数据缺乏实时性，无法进行实时预测，从而可能无法达到预期的干预结果。为此，本文基于学生在校园内学习、生活时产生的实时行为数据，结合问卷调查、人口统计学等相关的数据来进行成绩预测等相关的大数据研究。

成绩预测在教育管理中起到重要的作用。当前，挂科现象在大学生中非常普遍，甚至有人认为不挂科的大学生活是不完整的。然而，挂科可能会造成学生无法按时毕业或者无法找到心仪工作的后果。因而如果能提前发现学生的学习异常，通过引导和干预就有可能阻止这些不幸事情的发生。学习异常的发生可能源自于学习态度或者学习目标的转变，而这种转变是可以在学生的日常生活中表现出来的。大学校园本身就是一个小型的社会系统，其内部服务体系几乎可以满足学生绝大多数的需求。而校园服务的实现，如食堂吃饭、超市购物、图书馆借书、出入宿舍、教学楼打水等，大多数是通过校园“一卡通”来完成的。因而学生在校园中的食堂、超市、教学楼、宿舍楼、图书馆之间的日常生活轨迹就通过“一卡通”以数字化的形式记录下来。然而，大家并不知道行为和成绩之间的关系，也不知道行为变化和成绩变化之间的关系。

针对这种需求，基于这些“一卡通”记录下来的行为信息，特别设计了学生画像系统。该系统量化了心理学中影响学生成绩最重要的两个指标：努力程度和生活规律性，作为系统中的画像因子。努力程度包括去教学楼、图书馆消费的次数，对应到学生上自习或者上课的次数，反映了学生花在学习上的时间多少。而生活规律性包括出入宿舍的规律性、吃饭特别是

吃早饭的时间规律性、洗澡洗衣服的时间规律性、购物的规律性等，与学生的自我控制与自我约束能力密切相关。通过分析这些画像因子和成绩之间的关系，发现努力程度和生活规律性与成绩呈显著正相关性。特别地，针对某个年级的4年数据使用相关性计算，发现去图书馆的次数和成绩的序相关性达到0.3（$p<0.01$），而洗澡规律性和成绩的序相关性稍弱，为0.17（$p<0.01$）。图1展示了目前研究的所有努力程度和行为规律性的指标与成绩的相关性。更进一步地，分析行为变化和成绩变化之间的关系，发现努力程度和生活规律性的增加也会导致学习成绩的提升。因而，对于学习越努力、生活越规律的学生，他们的学习成绩越好。同时，基于学生在同一地点共现的次数，构建学生在校园内的社交关系网络，并分析每个学生的学习成绩和朋友间的学习成绩之间的关系。笔者发现，每个学生的成绩和朋友的平均成绩呈正相关的关系。这不仅验证了社会学中的成绩上的同质性，还能帮助构建准确率更高的成绩预测系统。

针对努力程度、生活规律性和社交关系网络以及过往的学习成绩，设计了多任务迁移学习算法来进行未来成绩的预测。该算法不仅通过多任务特性考虑了特征相关性存在学院之间的差异性，而且还通过迁移学习特点考虑了不同学期之间相关性的变化。同时，为了更好地保护学生隐私，将成绩变换成排名，并进行归一化，利用排序学习算法来进行学习。当学生的数据缺乏或者缺失时，该算法利用朋友的加权预测成绩作为学生的预测结果输出。在测试时，给定前5个学期的数据作为训练集，预测第6个学期的成绩排名，以预测排名和实际排名的序相关性作为预测算法性能评价的指标。最终预测算法的序相关性高达0.9，这让算法在实际中被广泛运用成为可能。而且，基于“一卡通”对于记录行为的实时性设计的成绩预测模型，可以帮助教育管理者及时发现学生的学习和生活异常情况，从而能对学生进行及时的干预和引导，从而实现从传统教育中的后置性应急到前置性预警引导的转变，实现从离线静态分析到自适应性地动态分析的转变。

除了发现这些行为数据在预测成绩方面起到的重要作用以外，笔者还发现了它们在贫困生检测、毕业去向预测、馆藏图书推荐等方面的价值。不同家庭经济条件的学生在消费行为方面可能呈现较大差异，因而消费行为数据对于贫困生检测存在一定的作用。学生毕业时的去向由很多原因决定，不仅取决于学生的成绩，还包括实习和科研等课外活动的经历、生活作息的规律性、家庭的经济状况等。而行为数据的存在给馆藏图书推荐带来较大变化，不仅可以区分男女生在借书上的差异性，也

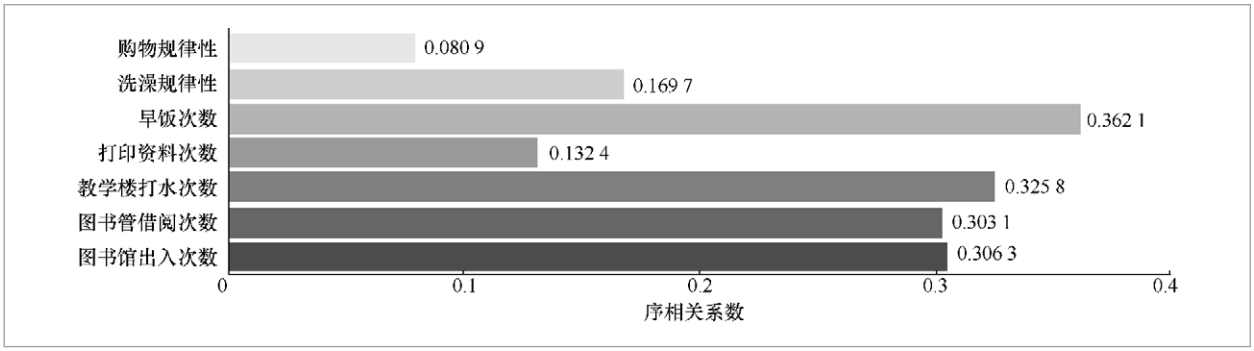


图1 行为规律性和努力程度与成绩的序相关性

可以区分成绩不同的学生在图书借阅上的偏好；反过来，通过学生借阅的图书信息，也能辅助确定学生的成绩信息。通过这些研究发现，当前的大学校园内，已经积存了很多对学校教育管理具有重要战略价值的数据。虽然对这些数据的价值已做了初步探讨，但是仍然还有待进一步的探索与发现。

参考文献

[1] Anderson J R, Boyle C F, Reiser B J. Intelligent tutoring systems. Science, 1985, 228(4698): 456~462

[2] Romero C, Ventura S. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2010, 40(6): 601~618

[3] Lindsey R V, Khajah M, Mozer M C. Automatic discovery of cognitive skills to improve the prediction of student learning. Advances in Neural Information Processing Systems, 2014:1386~1394

[4] Ritter S, Anderson J R, Koedinger K R, et al. Cognitive tutor: applied research in mathematics education. Psychonomic Bulletin & Review, 2007,14(2): 249~255

[5] Qiu J Z, Tang J, Liu T X, et al. Modeling and predicting learning behavior in MOOCs. Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM'16), San Francisco, USA, 2016 Accepted

[6] Ramesh A, Goldwasser D, Huang B, et al. Learning latent engagement patterns of students in online courses. Proceedings of the 28th AAAI Conference on Artificial Intelligence, Quebec City, Canada, 2014

[7] Anderson A, Huttenlocher D, Kleinberg J, et al. Engaging with massive online courses. Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 2014: 687~698

[8] Seo M, Hajishirzi H, Farhadi A, et al. Solving geometry problems: combining text and diagram interpretation. Proceedings of EMNLP, Lisbon, Portugal, 2015

[9] Hosseini M J, Hajishirzi H, Etzioni O, et al. Learning to solve arithmetic word problems with verb categorization. Proceedings of EMNLP, Doha, Qatar, 2014

[10] Tamhane A, Ikbal S, Sengupta B, et al. Predicting student risks through longitudinal analysis. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2014: 1544~1552 □