中国远程教育
DISTANCE EDUCATION IN CHINA

# 网络教学平台学生学习行为聚类分析*

□ 田 娜 陈明选

【摘 要】 随着数字化学习系统的使用和流行，学生在与系统交互的过程中产生了大量的原始数据。因此，数据挖掘技术可以用来从这些数据中提取出有用的信息以改进高等教育机构的管理、教学和研究效率。例如将聚类算法、决策树和关联规则方法应用到高等教育过程中，可以帮助改进学生的学习表现、辅助选择课程和学校补助基金的最优化管理等等。本文以江南大学网络教学平台为例，采用数据挖掘技术，根据学生的相似特性对学生进行聚类分析，以分析对课程成绩影响的各种因素。另外，还对学生在课程讨论区的活跃程度进行了社会网络分析。

【关键词】 教育数据挖掘；网络教学平台；分组模型；聚类分析

【中图分类号】 G40-057 【文献标识码】 A 【文章编号】 1009—458x（2014）11—0038—04

DOI:10.13541/j.cnki.chinade.2014.11.007

过去几十年以来，计算机的普遍使用推动了各种数字化学习系统在教与学过程中的应用。然而对系统的改进需求从未停止。学生在搜索学习资料的时候，会在系统中遗留下一系列的痕迹或者个体特征信息。因此，就有了一个很重要的研究问题：利用这些数据我们可以做些什么呢？当然，对这些数据的分析远远比单纯记录要复杂得多，教师可以利用其中的信息去改进教学以提高教学绩效[1]。数据挖掘技术作为一门成熟的技术，已经广泛应用于很多领域，比如医药、商业、市场调研等等，从海量数据中挖掘有用的信息，以辅助决策，有利于行业政策的制定。但教育数据挖掘相对来说是一个新兴的学科[2]。Romero等罗列了1995-2005期间数据挖掘在传统教育中的应用[3]。Maimon等对知识发现和数据挖掘在教育中的应用进行了详细的介绍，只不过均集中在对概念和理论的介绍，没有落实到真实数据中[4][5][6]。Romero等为了改善教与学的质量，应用不同的数据挖掘技术，将从学习管理系统中收集到的数据与传统的教学测试结果相结合，以确定网络学习模块是否适合学生，并且根据学生对系统的使用记录来将学生进行分类[7]。Delavari等分析了课程的测试结果，以确定决定学生取得优异成绩的关键因素[8]。研究者们还采用数据挖掘预测和发现了影响最后考试结果的不同因素之间的关系，从而有针对性地去调整教学以提高学习绩效。也有研究者采用分类（classification）方法来确定学生的最终成绩。也可以识别不正确使用系统的学生，以改善个人以及小组的成绩。Romero等指出数据的冗余和属性的相关性，从而增加了知识发现的难度[9]。为了体现反馈在在线测试中的重要性，作者在Moodle平台中进行了有73个学生参与的8个在线测试，其中包含了多项选择题。收集的数据包括：正确率，花费的时间，反馈要求，提供的反馈以及学生是否检查了反馈。李婷等经过对近年来国内外教育数据挖掘领域的大部分文献作了综述性研究，我们为了解当前国内外学者的研究动态给了很多指导和灵感[10]。在数字化学习系统（例如智能辅导系统）支持下的学习，由于具备了追踪学生学习过程的功能，比传统的面授辅导更有优势。并且，学生之间还有交流和协作，比独自学习效果更好，还可以锻炼学生的团队协作能力。因此，对学生的评价也可以将团队协作能力考虑进去。Moodle是一个被广泛采用的开源的网络学习管理系统，它可以跟踪并记录学生的每一个学习动作，故本文将在Moodle平台上应用数据挖掘技术以提取有用的知识。我们对学生的学习过程与学习结果

* 基金项目：本文受江南大学自主科研基金（项目编号：1245210382130120, 1242050205142810）资助，国家高技术研究发展计划（项目号：2013AA040405）。

进行记录与分析，应用聚类分析技术对学生进行分组，以及应用社会网络分析对学生在网络教学平台上的活跃程度进行分析，以确定网络教学平台对学生学习的辅助效果，并对其进行量化分析。

## 二、教育数据挖掘

顾名思义，教育数据挖掘就是数据挖掘技术在教育中的应用。数字化学习系统提供的数据可以转化为有用的信息，教师可以利用这些信息来改进教学过程；学生也可以从这些信息中受益，比如获取合适的学习内容，调整自己的学习过程等等。因此，教育数据挖掘的目标是深入了解教与学，从而更好地改进教与学的质量。在传统的教室里，老师观察学生的行为并且分析测试结果，根据学生的反馈调整教学。每个学生根据自身的特征对老师的调整有所回应，如此循环。然而，在数字化学习系统中，这样的交互反馈也会有一定程度上的丢失，所以要利用数据挖掘技术来辅助调整。

1. 教育数据挖掘的流程

从数字化学习系统中收集到的数据包含了个体的特征数据（学生、教师、管理人员）和学术数据（学习过程）。那么数据挖掘技术可以用来安排教学进度，预测学生的学习行为（辍学原因等）。通常来讲，教育数据挖掘过程包含了四个阶段：数据收集、数据预处理、数据分析和结果解释。整个过程是迭代的、不断完善的，具体流程图如图1所示，在对结果进行解释之后，采用挖掘到的知识来调整教学过程，然后新的一轮迭代过程又开始。当然，调整对教学效果的影响可能是正面的也可能是负面的。

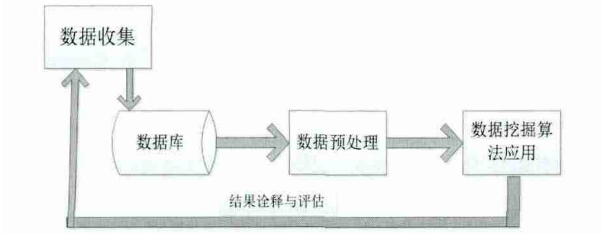


图1 数据挖掘流程图

2. 分析模型的分类

数据挖掘分析模型可以分为两类：描述性模型和预测性模型。描述性模型主要用在统计与可视化中以辅助决策制定；而预测性模型主要用在数据预测中，比如学生的分数，通过率等等。本文选择了数据挖掘技术中的聚类分析作为描述性模型进行案例分析。聚类分析是将数据集划分为具有相似性质的类。聚类算法不需要在事前知道有几个类别，这是与数据分类分析最大的区别。本文选择基于欧几里德距离的K-means算法。K-means算法将数据集划分成K个类，每个类均有一个中心点，中心点是由类内所有的点取算数平均值得到。而初始类中心点的位置是随机产生的。将学生进行分组，从而可以预测和分析学生对不同教学策略的反应，以及识别出错误使用系统以及玩游戏的学生，从而进行特殊辅导。如下图2所示：

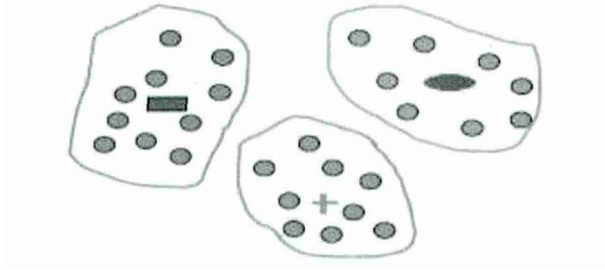

图2 根据学生的学习特征将学生分类(Clusters)

聚类的目标是使得簇（类）内距离最短，簇（类）间距离最大，用以下公式作为目标函数：

$$f=\sum_{i=1}^{K}\sum_{x_j\in S_i}\left(x_j-\mu_i\right)^2 \qquad (1)$$

算法过程如下：

从条记录中随机选取一条记录作为簇的中心；

对剩余每条记录测量其道每个类中心的距离，并把它归到最近的类；

根据归到每个簇的记录重新计算类中心；

迭代（2）和（3）直至公式（1）满足指定阀值。

K-means聚类算法的优点是速度快，容易根据聚类的结果对类的数量进行重新调整。

3. 教育数据挖掘工具

已经被研究者广泛运用的，以及还具有巨大探索空间的学习分析技术有统计与可视化、聚类、预测、关联规则挖掘、社会网络分析、话语分析和内容分析。七类分析技术各有所长，要了解学习现状的全貌往往需要多种分析方法和工具的整合。新西兰怀卡托大学开发的怀卡托智能分析环境（Waikato Environment for Knowledge Analysis，Weka）、Microsoft Server Analysis Services（SSAS）、专业统计软件SPSS均能实现聚类算法。

社会网络分析对社群间的交互行为予以考察，对个体在群体的中心度、个体间相互之间的距离以及大

群体中的小团体进行挖掘和呈现。一些主流的社会网络分析工具，包括美国加州大学欧文分校（UC Irvine）人员编写的Ucinet和澳大利亚卧龙岗（Wollongong）大学开发的针对Moodle、Blackboard等主流教学系统的SNAPP和微软研究院研究人员等研发作为Excel插件使用的NodeXL。该分析方法也存在自己的短板，一是“重频次轻深度”，以个体间的交互频次作为“远近亲疏”的分析依据，对交互内容的浅显深刻却未有甄别；二是“重结果轻过程”，往往截取某一时间点的社会网络现状作为分析对象，难以获得整个网络形成和发展的历程。本文是使用Weka和Ucinet来完成聚类和社会网络分析。

## 三、案例分析

我们所需要的数据是从人文学院的网络教学平台上收集得到，研究样本包括了选修“程序设计语言C”的65个学生。第一步，学生先做一个学前测试以确定他们的初始知识状态。第二步，学生用网络教学平台学习一个学期之后，再做一个学后测试。为了获得所需要的数据（比如在课程和活动中花费的时间，访问频率等），作者使用不同的SQL查询。教学平台的日志文件还详细地记录了用户的每一步动作。课程信息在数据库中的表示如下表1所示：

表1 课程表格

| Table | Contains |
|---|---|
| mdl_lesson | Lesson settings |
| mdl_lesson_pages | Lesson page content |
| mdl_lesson_timer | Start and end lesson time |
| mdl_lesson_grades | Students’ grades |
| mdl_lesson_branch | Access information (content) |
| mdl_lesson_attempts | Access information (questions) |

网络教学平台中有超过200个表格，这里就不再一一列出。

1. 聚类与相关性分析

我们的目标是根据系统的日志文件和测试结果，将具有相似特征的学生分成一组，以辅助教师制定个性化的教学和辅导计划。网络教学平台存储了关于学习过程和个体信息的所有数据。利用MySQL将查询得到的结果导出为ARFF格式的文件，用Weka中的K-means算法进行了聚类分析。结果是将学生分成了两类，一类是包含了学前测试和学后测试成绩优秀的学生；第二类包含了网络学习比较活跃和花费时间比较多的学生（见表2）。

表2 每个类的中心点

| 指标 | Cluster 1 | Cluster 2 |
|---|---|---|
| 前测 | 44, 95 | 37, 06 |
| 后测 | 71, 91 | 48, 27 |
| 活跃程度 | 222, 04 | 383, 20 |
| 在线时间 | 116, 72 | 177, 53 |

考虑到某些指标对测试结果影响，本文做了相关性分析。相关性分析是统计学中用来测量两个变量之间关系，数值在“-1”到“1”之间。关联度为“0”表示变量之间没有关系。负关联度意味着一个变量变大，另外一个变小；正关联度意味着两个变量在同一个方向上变化；如下表格所示：

表3 关联性结果样例

| Variable 1 | Variable 2 | Correlation |
|---|---|---|
| 完成的课时 | 课程成绩 | 0.37 |
| 作业分数 | 前测成绩 | 0.46 |
| 作业分数 | 后测成绩 | 0.40 |
| 教学材料的浏览次数 | 课程成绩 | 0.54 |
| 登录次数 | 后测成绩 | -0.39 |
| 页面停留时间 | 后测成绩 | -0.42 |

由上表可以看出，完成的课程数目与课程的分数是呈正相关性的。课程最终成绩优秀的同学也在学前测试和学后测试中获得了好成绩，这说明这些学生在学前和学后都是优秀的。访问过的课程页面的数量与课程分数也是呈正相关性的。然而，课程的浏览次数与在网络平台上停留的时间却与学后测试成绩呈负相关性，产生此结果的原因可能是学生在登录网络平台之后，受了学习之外活动的干扰，比如网络游戏或者实时通讯工具聊天以及社交媒体网络等等。

2. 社会网络分析

除了从学前测试成绩、学后测试成绩、学习资料浏览次数以及下载次数等各个方面考虑主子之外，还考虑学生在网络教学平台上与教师和同学之间的互动，并对发帖和回复帖的次数进行了统计。

该门课程教学材料、课程通知两模块主要发生“教师发布、学生察看”的单向行为，师生双向交互行为集中在课程讨论区、课程问卷以及个人答疑三个模块。其中讨论区模块学生操作最为频繁，且学生之间也可以直接产生交互行为。统计分析该模块的师生行为，更能准确反映社会网络的现实状况。

讨论区模块师生的具体操作行为有读帖、发帖、回帖三种，回帖作为两个体间交互的主要表征，对其进行统计。为适应分析工具Ucinet的输入数据要求，预处理数据形成关系矩阵，表4显示了部分师生的交互频次。其中行与列分别对应被回帖师生个体编

号和回帖师生个体编号，表中数值为回帖个体回复被回帖个体的帖子数。

表4 师生交互矩阵

| i / j | 6 | 204 | 207 | 208 | 209 |
|---|---|---|---|---|---|
| 6 | - | 1 | 0 | 0 | 0 |
| 204 | 1 | - | 0 | 0 | 0 |
| 207 | 3 | 0 | - | 1 | 0 |
| 208 | 7 | 0 | 1 | - | 1 |

分析发现该课程存在两种成分，其中一种成分共28名成员，包括1名任课教师和27名学生，该成分中的成员在课程讨论区与其它课程成员产生过至少一次交互（回帖或被回帖）；另一种成分是课程其余的37名学生，课程从始至终未与他人在讨论区进行交互（未交互并不代表在讨论区的活跃度为“0”，存在发过首帖但无人回帖，也未回别人帖子的情况）。图3显示了第一种成分所有成员的相互联结情况。其中箭头起始点为回帖者，指向点为被回帖者；浅色结点表示回帖者，深色结点表示被回帖者（同时有回帖和被回帖行为的结点标记为深灰色）。

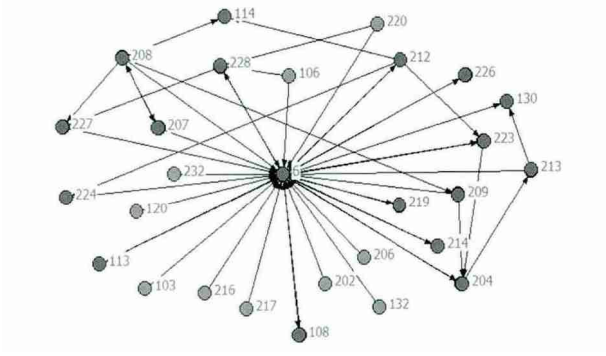


图3 师生课程讨论区交互网络

联系电子档案中最终课程成绩，是否产生交互行为对成绩的影响显著，产生交互行为的成分所有成员平均得分76.8分，中位分数77分；未产生交互行为的学生平均72.5分，中位分数70分。即产生交互行为的学生趋向于取得较为理想的成绩。仅考查产生交互行为的成分，鉴于一种中心性项目中所有学生的数值较为相近，区分度不大，最终得到“显著性大于0.05”的结果，即学生在社交中扮演不同角色与所获成绩间没有明显关系。

总之，本文在从网络教学平台中收集得来的数据上应用了数据挖掘技术，本文的主要工作是应用聚类分析来对学生进行分组，学前测试与学后测试成绩之差用来衡量学生是否有进步。原本期望的结果是，在网络平台上比较活跃，花时间比较多的同学成绩应该会比较好，但是分析的结果却是相反的，这说明成绩还与其他因素相关。比如有的同学没有电脑或者网络，他们的网络学习集中在实验室，所以学习效率会比较高。而有些同学可能在网络平台课程上登录并停留了很长时间，但是效率会比较低，可能学习过程被其他的活动所干扰。而且聚类分析得到的结果与相关性分析类似，得分比较低的同学在网络教学平台上比较活跃，故被归为一类。进而教师可以有针对性的调整整个教学过程，以帮助学生更好的获得知识，避免一些可能引起的负面影响。此外，本文还对学生在课程讨论区的地活跃程度进行了社会网络分析，分析结果显示学生在社交中扮演不同角色与所获成绩间没有明显关系。

[参考文献]

[1] R. Llorente, M. Morant. Data Mining in Higher Education [J]. New Fundamental Technologies in Data Mining, 2011, 201-220.

[2] V. Kumar, A. Chadha. An empirical study of the applications of data mining techniques in higher education [J]. International Journal of Advanced Computer Science and Applications, 2, 2011.

[3] C. Romero, S. Ventura. Educational data mining: A survey from 1995 to 2005 [J]. Expert Systems with Applications, 2007, 33: 135-146.

[4] O. Maimon, L. Rokach. Introduction to knowledge discovery and data mining [M]. Data Mining and Knowledge Discovery Handbook, New York: Springer, 2010, 1-15.

[5] L. Rokach, A survey of clustering algorithms [M]. Data Mining and Knowledge Discovery Handbook, New York: Springer, 2010, 269-298.

[6] R. Baker, Data mining for education [M], International Encyclopedia of Education (3rd edition), Oxford, UK, Elsevier, 2010, 112-118.

[7] C. Romero, S. Ventura et. al., Data mining algorithms to classify students [C]. Proceeding of Educational Data Mining, 2008, 20-21.

[8] N. Delavari, P. A. Somnuk, Data mining application in higher learning institutions [J]. Informatics in Education, 2008, 7:31-54.

[9] C. Romero, S. Ventura, E. Garcia, Data mining in course management systems: Moodle case study and tutorial [J]. Computers & Education, 2008, 51 (1): 368-384.

[10] 李婷，傅钢善. 国内外教育数据挖掘研究现状及趋势分析[J]. 现代教育技术，2010，10（20）：21-25.

收稿日期：2014-07-30

作者简介：田娜，副教授，博士；陈明选，教授，院长。江南大学教育技术系（214122）。

责任编辑 碧 荷