MOOC 学习行为分析及成绩预测方法研究

文/郝巧龙　魏振钢　林喜军

摘要

近年来，MOOC 以“互联网+教育”模式迅速发展，积累了海量学习行为数据，MOOC 学习行为分析及成绩预测成为研究热点。笔者收集学习行为数据，用 Clementine 构建了 MOOC 成绩预测模型，为验证其有效性，依托智慧树平台数据结构课程的行为数据展开实证研究，旨在为其课程团队提供指导意见。

【关键词】MOOC　数据挖掘　回归分析　成绩预测模型

MOOC(Massive Open Online Course) 的理想是任何人在任何时间和地点学到任何知识。2012 年斯坦福大学等名校组建了 Coursera、Udacity 和 edX 平台。2013 年清华北大等名校和互联网公司展开了 MOOC 实践，研发了学习者在线交互平台，为分析成绩与行为的关系提供数据支持。国内在部分课程上进行 MOOC 教学但实证研究较少。蒋卓轩 [2] 首次描述中文 MOOC 学习行为并预测成绩。Suhang Jiang 用绩效考核和公开课结合进行一周的干预，用 logistic 回归分析预测成绩验证了及时干预的激励作用。笔者理论上对比国内外学习行为分析及成绩预测成果，用线性回归分析构建了 MOOC 成绩预测模型；实践上用 Clementine 进行实证研究，预测效果良好并提出应用方案，为教师的决策支持提供严谨的数据保障。

1 成绩预测模型构建

笔者分五个模块构建了 MOOC 成绩预测模型（图 1）。

模块一：确定变量初始集。根据预测目标确定变量范围，回归分析的前提是因变量为数值型变量。

模块二：全部变量进入回归方程。选择进入法和逐步法将全部变量加入方程中便于对比预测结果。

模块三：筛选重要影响变量进入回归方程。为保证结果的普适性，需要对变量初始集依次进行散点图分析、统计量分析和特征选择，剔除相关性弱的构成变量集 3。在变量集 3 上选择上述两种方法进入方程。

模块四：评估预测模型确定最优回归模型。方程通过回归方程的拟合优度检验、回归方程的显著性检验和回归系数的显著性检验后才能用于实际问题，通过评估模块二、三的模型确定最优模型。模块五：呈现预测模型结果。模型结果直观呈现变量关系，形式为 $Y=\varepsilon+\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k$。因变量 Y 为总成绩；$\varepsilon$ 为误差且 $\varepsilon \sim N(0,\sigma^2)$；$\beta_0\cdots\beta_k$ 为未知参数；自变量 $X_1\cdots X_k$ 为影响因素。

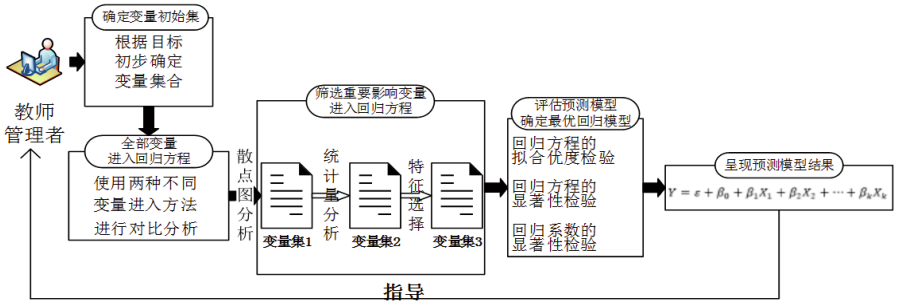

图 1：成绩预测模型

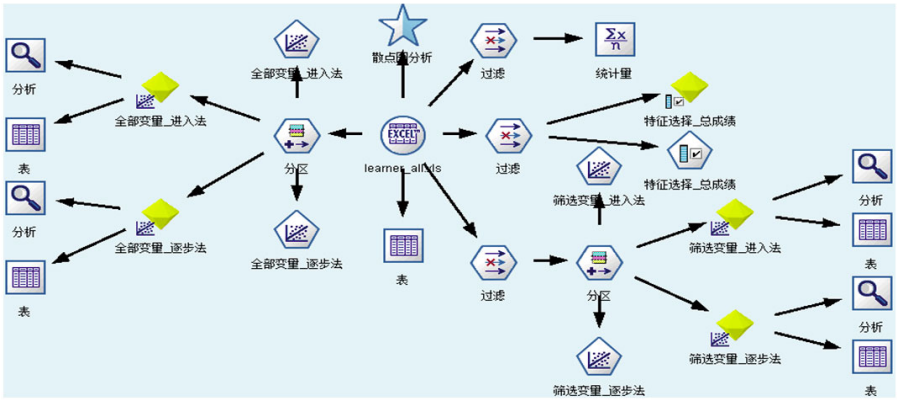

图 2：模块二、三数据流

2 实证研究

在 2015 年 3 月至 6 月数据结构的学习行为数据上展开研究，因为有本校学生参与，分析结果可信度强。所用设备是 Intel Core i3 处理器，Win 7 操作系统，2.27GHz 主频，6G 内存。

2.1 研究过程

模块一：预处理行为数据存入 learner_all 表得到变量初始集。总成绩为因变量 Y，自变量是学生编号、学校编号、持续时间、学习进度、观看时长、笔记数、一～七次作业成绩、发帖数、回帖数、得分帖数、一～六次见面课成绩、在线成绩、论坛得分、见面课成绩和期末成绩，编号为 X_1~X_{26}。

模块二：X_1 不起作用将其过滤；将总体样本分区 70% 为训练集 30% 为测试集；添加回归模型目标为总成绩，选择进入法和逐步法使全部变量进入方程，为模型结果连接表输出和分析节点，执行数据流（图 2）显示结果。

模块三：一是散点图分析：读入 learner_all 设置总成绩为输出其它变量为输入；添加散点图节点 Y 轴为总成绩 X 轴为 25 个自变量，X_2、X_9、X_{10} 与 Y 不相关将其剔除得到变量集 1。二是统计量分析：添加统计量节点由 Pearson 相关性强度得知 X_3 与 Y 相关性弱剔除后构成变量集 2。三是特征选择：添加特征选择节点目标为总成绩输入为 21 个自变量，其中 X_4 变异系数低将其剔除形成变量集 3。为变量集 3 添加分区节点设置同模块二；添加回归模型选择进入法和逐步法执行，为模型结果连接表输出和分析节点，执行数据流（图 2）显示结果。


表（31 个字段，227 条记录）#2

文件　编辑　生成

	总成绩	分区	$E-总成绩	$E1-总成绩	$E2-总成绩	$E3-总成绩
1	93.0...	1_训练	92.525	92.576	92.648	92.648
2	98.0...	1_训练	97.825	97.866	97.830	97.830
3	100....	1_训练	100.163	100.077	100.013	100.013
4	55.0...	2_测试	54.572	54.516	54.721	54.721
5	97.0...	1_训练	96.603	96.577	96.614	96.614
6	94.0...	1_训练	94.141	93.972	94.029	94.029
7	77.0...	1_训练	76.824	76.859	77.064	77.064
8	92.0...	1_训练	92.088	92.116	92.086	92.086
9	98.0...	1_训练	97.973	97.897	97.861	97.861
10	93.0...	1_训练	93.024	93.074	93.002	93.002
11	54.0...	1_训练	54.045	54.049	54.075	54.075
12	76.0...	1_训练	75.955	75.966	75.972	75.972
13	96.0...	2_测试	95.880	95.913	95.966	95.966
14	93.0...	1_训练	92.595	92.566	92.606	92.606
15	94.0...	2_测试	94.009	93.960	94.003	94.003

图 3：输出表节点结果

模块四：

（1）回归方程的拟合优度检验

依据判定系数 R^2 和估计标准差来检验，

<< 下转 168 页

基于主题网络爬虫的信息数据采集方法的研究与应用

文/盛亚如　魏振钢　刘蒙

摘要

互联网上的各种信息以数百万级的方式增长着，而这些信息又大多是散乱分布的，无法满足人们所要求的整合信息分析的需求，传统的采集和收集方法又很难满足要求。因此本文提出利用主题网络爬虫的概念和方法，运用正则表达式去匹配出网页中所需要的特定信息数据，有效的增强爬虫程序的适用性、缩短用户获取信息的时间。并将此方法应用于二手房信息数据采集中，包括价格、户型、楼层等基本数据，建立起了一个统一的二手房数据库。

【关键词】主题网络爬虫 正则表达式 二手房

1 引言

互联网上的信息数据以爆炸式的方式增长着，而这些信息数据内容又大多是基于页面形式的，其中包含一些非结构化的数据，如文字、图像、视频等。如果只是采用人工化的方式对信息数据进行采集，已经很难满足人们的要求了。因此有必要采用某种技术或手段从互联网上自动采集信息数据。

网络爬虫能实现对互联网信息数据的自动采集，从而弥补了人工采集的缺陷。网络爬虫是随着搜索引擎发展而产生的一种通用信息采集技术，是搜索引擎中的核心部分，它根据用户要求从互联网上下载网页，尽可能多的抓取网页中的相关链接和内容，并能沿着链接继续爬行，是一种能力强大的信息采集程序。

2 主题网络爬虫

主题网络爬虫是在通用网络爬虫的基础上进行的延伸，根据某一领域内特定的主题进行相关信息的查询，搜索互联网抓取下载网页，从网页中采集相关信息数据和超链接。它并不会访问所有的网页，而是在访问前就判断超链接、锚文本、文本等与主题的相关度，按照相关度的高低来决定访问的优先级顺序。

主题网络爬虫的主要思想就是：把用户搜索的查询词作为主题，从选定的初始 URL 出发，访问网页中的所有超链接，根据某种搜索策略对这些 URL 进行主题相关度预测，将符合要求的 URL 加入待访问队列中，并按照某种优先级排序从队列中抽取 URL 来作为下一次要访问的对象，按照这种规律执行下去，直到待访问队列为空或者满足某种停止条件为止。

3 基于主题网络爬虫的信息数据采集方法与应用

通过分析网站页面时发现，页面中关于某一项主题的结构和框架都是一样的，因此可以考虑运用正则表达式去匹配出页面中我们所需要的链接和内容。下面以安居客网站为例进行二手房数据的采集。

3.1 网站页面分析

<< 上接 167 页

R^2 越接近 1 表明拟合优度越高。进入法使变量进入方程 (无论筛选变量与否)，R^2 均为 1 表明拟合优度高。逐步法进入方程 R^2 为 1 估计标准差为 0.314 小于进入法的 0.331，显示出逐步法的优越性且拟合优度提高。

（2）回归方程的显著性检验

依据概率 p 值、残差平方和、残差均方进行检验，p 小于 0.05 表明因变量与所有自变量线性关系显著。进入法使变量进入方程 (无论筛选变量与否)，p 为 0 线性关系显著。表明筛选变量后方程变精练预测能力未减弱。逐步法建模后残差均方减至 0.110 小于进入法的 0.111，p 为 0 线性关系显著。

（3）回归系数的显著性检验

依据概率 p 值进行检验，p 小于 0.05 表明自变量与因变量线性关系显著。进入法使全部变量进入方程，仅 7 个变量 p 值小于 0.05 线性关系不显著。进入法使重要影响变量进入方程，较多变量 p 值大于 0.05 但值变小。表明筛选变量后线性关系有改善。逐步法建模 p 最大为 0.02 表明线性关系显著。

为直观展示预测效果，连接四个回归模型添加分析、评估和输出表节点。全部变量 _ 进入法对应为 $E- 总成绩，筛选变量 _ 进入法对应 $E1- 总成绩，全部变量 _ 逐步法对应 $E2- 总成绩，对应 $E3- 总成绩。分析节点结果表明测试集的最大 / 小误差比训练集小，且 $E3- 总成绩最佳。评估节点结果显示 $E3- 总成绩增益明显接近最佳线。图 3 展示了训练集和测试集的预测值与总成绩吻合。综上所述，最优回归模型是筛选变量 _ 逐步法所得的模型。

模块五：结果表达式为

$$Y = -0.0058 - X_{14} * 0.018 - X_{15} * 0.034 - X_{16} * 0.118 + X_{23} * 0.999 + X_{24} * 1.046 + X_{25} * 0.997 + X_{26} * 0.993 + X_{21} * 0.073$$

2.2 研究结果

2.2.1 结果分析

结果表明系数不同对总成绩的影响也不同。X_{24}、X_{23}、X_{25} 和 X_{26} 权重较大。论坛中发 / 回帖数反映学习积极性，得分帖数反映知识掌握程度，论坛参与越积极总成绩越高；在线学习时观看视频次数越多知识掌握越牢固，自主学习能力越强越及时提交作业；见面课是学习者与教师进行互动探讨极大提升积极性；梳理前期知识能显著提高期末成绩。

2.2.2 应用方案

一是学习者进行自我干预；二是教师和管理者对学习者进行人工干预；三是开发者接受学习者的建议后对学习者进行系统干预。

学习者应对重点环节做出自我调整，提高自主学习能力，缩短学习懈怠时间。教师和管理者应精心设计教学视频和题库，激发学习兴趣提高在线成绩；论坛讨论应缩短答疑时间，高质量帖子应加分；见面课是人工干预的好时机，能直观地调动各校学习者的积极性，及时解决疑难点；期末考试题的设计应有区分度。开发者应以改进在线体验和提供优质资源为目标，增加个性化制定学习计划模块，根据学习者设置的自我干预条件及时提醒和系统干预。

3 结束语

笔者宏观上运用多元线性回归分析构建了普适的成绩预测模型，微观上进行实证研究，所得表达式使得教师和学习者可直接定位重点模块，同步提高教和学的效果。预测结果为教师和管理者的决策支持提供了严谨的数据保障，为后续学习行为分析及成绩预测起到借鉴和推动作用。

参考文献

[1] 汤敏．慕课革命：互联网如何变革教育 [M]. 北京：中信出版社 ,2015.

[2] 蒋卓轩，张岩，李晓明．基于 MOOC 数据的学习行为分析与预测 [J]. 计算机研究与发展 ,2015,03:614-628.

作者单位

中国海洋大学信息科学与工程学院　山东省青岛市　266100