文章编号：2096-1472(2020)-01-24-03
DOI:10.19644/j.cnki.issn2096-1472.2020.01.006

大数据背景下在线学习数据分析方案设计

程 香[1]，程长征[2]

(1.中共安徽省委党校信息技术中心，安徽 合肥 230022；
2.合肥工业大学土木与水利工程学院，安徽 合肥 230009)

摘 要：大数据环境下数据分析是数据价值发掘的重要过程，合理的数据分析方案对数据分析过程和成效至关重要，提出一个大数据分析流程方案设计方法。该方案详细给出了分析流程应包括的内容，以及每个阶段的主要任务和结果形式，从而能有效指导数据分析项目的执行。该方法成功地应用于在线学习课程成绩预测分析项目中，对其他数据分析项目也具有通用性。

关键词：大数据分析；分析方案；分析需求
中图分类号：TP39 **文献标识码**：A

A Scheme of Online Learning Data Analysis Under the Big Data Background

CHENG Xiang[1],CHENG Changzheng[2]

(1.*Information Technology Centre,Party School of Anhui Provincial Committee of C.P.C,Hefei* 230022,*China*;
2.*Department of Engineering Mechanics,Hefei University of Technology,Hefei* 230009,*China*)

Abstract: Data analysis is an important process for data value discovery in big data environment.A reasonable data analysis solution is crucial to the data analysis process and effectiveness.A method of big data analysis process scheme is proposed in this paper.The content that should be included in the analysis process,as well as the main tasks and results of each stage,are displayed in detail by the scheme,so as to guide the implementation of big data analysis projects effectively. This method was successfully applied to academic record prediction analysis project of some online learning courses with a public data set.It is also versatile for other data analysis projects.

Keywords:big data analysis;analysis scheme;analysis requirement

1 引言(Introduction)

在大数据背景下，海量数据的应用价值已经显现出来，进行数据分析与挖掘的研究与应用，对各行业都具有重要的战略意义。随着信息化程度的不断提升，教育行业内部信息的不断完善[1]，在线学习数据分析必然要上升到大规模级别。在线学习数据分析将成为解学习者过去和现在的学习状态，以及预测未来学习结果的一个重要手段[2]。虽然众多研究表明通过分析在线学习数据，向学习者和教师提供反馈建议，能够正向影响学习者行为和学习成绩[3]，但是对数据分析的实施过程关注较少。如何有效地进行在线学习相关数据分析和预测，已成为在线学习数据价值发掘的关键问题之一。

电子阅读材料有助提升学习者对认知概念的理解，在线测试能够实现学习者成绩自动化评估，因而在线学习颇受青睐。但由于在线学习是由学习材料、评估任务和通信媒介共同支撑，因此，待分析的数据不但来源广而且容量大，导致数据结构类型多[4]，数据分析难度提升。目前在线学习的研究集中于学习效果的各种影响因素，某种特殊行为及风险预测[5,6]方面，而这些研究实施过程的可操作性和分析结论的准确性，有赖于合理组织数据分析流程，详细规定各个阶段的主要任务，制定出尽量科学规范的大数据分析方案。本文结合在线学习课程公开数据集的学习成绩预测分析项目，详细介绍数据分析方案制定方法和过程，并且给出数据分析可视化结果。

2 大数据分析流程的方案(The scheme of the big data analysis process)

数据分析本质是在数据资料采集和整理基础上，运用统计和挖掘等科学分析方法，从数据中寻找客观事物发展特征和规律，从而得出有指导意义的结论以预测指导未来实践。

数据分析一般是对特定的领域问题展开研究，本文数据分析方案涵盖以下方面内容：首先要对其背景进行全面了解，从而更好地把握分析目标和方向；分析和思考用户需求，确定数据源的出处和所有可能的数据类型；拟定分析问题的难易程度及结果可能的呈现方式，搭建分析环境选择分析平台；根据分析目的确定算法建立模型，选择模型参数；评估分析结果判断模型是否需要改进，择优选用；其他数据分析应当考虑的内容。

基金项目：安徽省教育厅自然科学研究重点项目(KJ2017A939).

3　分析方案详细制定过程(Detailed analysis of scheme-making process)

3.1　问题定义

数据分析是利用数据统计和挖掘原理，从数据中获取知识和信息。第一步要进行领域和问题定义，尽可能详尽地了解所要分析项目的领域背景和挖掘目标，明确项目分析所要解决的问题。

edX是一个提供大规模开放在线课程的平台，旨在加强校园教育，推进教育研究，并增加全球在线学习机会。随着平台上课程持续上线，学习者数量不断增加，但课程低通过率和高退出率的状况令人堪忧。针对通过率问题，本文以学习者在线学习成绩预测为分析目标，识别处于成绩合格边缘的学习者，以便给予干预和支持。

3.2　分析工具选择

目前用于数据分析的软件众多，从价格、界面友好性和编程难易程度方面来衡量，R不失为一个很好的选择。R是开源软件包，也是一种编程语言，具有强大的数据分析和可视化能力[7]。

Hadoop采用并行计算且节点易扩展，可以提高算法执行的时间效率，并且能够解决内存等系统资源限制问题。对于涉及数据量大的分析项目，集成R和Hadoop作为数据分析平台，可以借助R丰富的组件库，拥有强大的数据统计分析能力，并且可以发挥Hadoop在分布式存储和计算方面的优势，进行全量数据分析。

3.3　数据的获取与处理

3.3.1　设计数据需求

为了对特定问题进行分析，需要相关领域的数据。

(1)数据源确定。在确定领域和即将分析的问题后，数据的来源就能够确定。根据挖掘分析问题的目标确定是否需要从Excel、其他统计软件、数据库、网页等载体中获得原始数据。

(2)数据获取。数据分析一般需要加载外部数据源，根据具体需求决定加载何种结构的数据源。对于CSV、TXT、Excel数据和网页中的数据，以及SPSS、SAS、Stata等统计软件中的数据，R均可使用其相应包中的函数读取；对于各种数据库中的数据，可以用自定义的函数连接到数据库，也可使用CRAN中多个连接数据库的R包，将各种系统中的数据载入R中。

案例从edx在线学习平台公开数据集获取分析数据，该数据集是2013学年上线的16门课程的学习者学习记录，共60多万条。本文通过函数加载学习者数据，为后续探查和处理数据，学习行为数据建模等数据分析过程准备数据。

3.3.2　探索性分析

对数据特征和规律进行初步探查研究，有助于后续分析过程决定采用何种预处理方式，以及选择何种算法和模型。对数据进行探索性分析可以从以下方式入手：

(1)数值指标。使用相关函数得到数据集的数字指标，了解数据的整体结构、变量情况、分布指标、缺失值等情况。该方法可以给出各项统计指标的确切值，有助于制作和观察图形、设定算法参数。

(2)可视化视图。图形显示数据将比数字化统计方法表现得更加形象生动和直观。对描述性和预测性分析数据结果，用R可将原始模型表达成丰富多彩的图形和可视化视图。

利用上述原理，案例分析项目结合数字化指标与可视化图形，对学习者行为样本数据进行探索性分析，以下给出部分探索性分析结果。

①分布情况。通过各行为变量的直方图可以了解相关变量的大致分布情况，案例数据集观测值的大多变量分布在最小的分组段，说明学习者数量虽然可观，但多数人学习投入程度不高。其中变量nchapters取值主要集中在0至2和2至4组段，可见多数人学习课程不足5个章节。

②缺失值。抽取学习者的基本信息、成绩和几项学习行为特征数据共11个属性，分析缺失值在数据集中的分布情况，发现涵盖所抽取属性的完整记录共有132751条，其余记录均存在1至8个不等的缺失值。

③相关性。对抽取的学习者学习行为特征向量进行相关性分析，结果如图1所示。从图中可见成绩与nevents、ndays_act、nchapters的相关性都比较高。

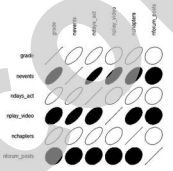

图1 成绩及学习行为特征向量相关图

Fig.1 Correlation diagram of grade and learning behavior trait vectors

3.3.3　数据预处理

为了改善数据质量，在将数据提交给算法和工具之前，需要对数据进行清洗、聚合、去噪、整理和格式化等一系列操作，并处理成适合进行挖掘分析的形式。采用何种预处理取决于数据探索性分析阶段了解的信息，这些信息包括相关变量的取值范围、缺失值情况、有无偏差及偏差程度等。

案例进行的数据预处理主要包括：在数据清洗阶段，去除内部不一致数据记录，并删除有缺失属性的记录。在数据归约阶段，对样本数据进行属性归约，选择影响学习成绩的学习特征行为向量作为学习者学习行为特征数据。在数据变换阶段，构造年龄属性，以便对不同年龄段的学习行为和成绩进行分组分析；此外还对成绩属性进行了离散化，将处于子区间[0,0.5]、(0.5,0.7]、(0.7,1]的成绩分别映射到1、2、3三个值；最后对特征向量进行标准化用于后续建模。

3.4　模型构建与评估

3.4.1　模型选择与构建

(1)模型选择

大数据分析需借助模型和算法进行挖掘操作，从数据集中发现有用的知识或模式。目前大数据分析采用比较多的挖掘算法是关联分析、聚类分析、决策树、分类与预测等[8]，对具体分析业务显然不可能使用固定的算法和参数来建立模型。并且算法模型应在机器学习的各阶段进行持续评估和改善，产生更好的分析效果。

对于算法和模型的选择，首先要确定使用何种算法及相应参数的大概取值范围。算法的选择主要由建模在数据挖掘应用中归属何种类别决定。对于参数范围的选择，采用试探法或者从数据探索性分析的结果来确定大致的范围。案例采用CART决策树和神经网络预测学习者成绩，并从中选择预测效果最好的模型。

(2)模型构建

构建神经网络模型(Artificial Neural Networks，ANNs)

和CART决策树(Classification and Regression Tree，CART)时将采用公共数据集中反映学生学习行为的5个属性作为自变量。随机选择3/4作为训练样本，剩下的作为测试样本。

对于神经网络模型，设定神经网络的输入节点数为5，输出节点数为1，隐藏层中的节点个数为6，权重值的衰减精度为0.0005。采用训练样本建模的分类正确率为95.14%，其混淆矩阵如表1所示。对于CART决策树模型，表1显示了利用训练样本建立的CART决策树模型混淆矩阵，该模型分类正确率为95.13%。

表1 训练样本集下神经网络和CART决策树的混淆矩阵

Tab.1 Confusion matrix of neural network and CART decision tree under training sample

模型	类别	预测值		
		1	2	3
神经网络	1	89939	0	1508
	2	787	1	1239
	3	1302	0	4785
CART决策树	1	90322	0	1125
	2	933	1421	1093
	3	1699	0	4388

(3)模型评估与使用

对模型进行评估目的是判断每个模型的预测能力，从而在多个模型中选择一个最优模型。可采用多种方法对模型的各方面进行测评，这些方法有：a.采用模型的混淆矩阵，讨论模型的预测结果和真实结果之间的差距；b.利用风险图，对模型的预测结果与真实结果之间的差别进行比较分析；c.通过绘制ROC图像进行模型评估；d.得分数据集。

案例利用测试样本对两个模型进行评价，评估模型的预测性能。表2为两种模型在测试样本下的混淆矩阵，模型预测的准确率分别为94.92%和94.95%。尽管两类模型的分类准确率都很高，但是预测值为2的类别的召回率TP/(TP+FN)显然均为0，无法正确标记出成绩处于及格边缘的学生类别。

表2 神经网络和CART决策树在测试样本下的混淆矩阵

Tab.2 Confusion matrix of neural network and CART decision tree under test sample

模型	类别	预测值		
		1	2	3
神经网络	1	29948	0	543
	2	239	0	422
	3	481	0	1554
CART决策树	1	30098	0	393
	2	284	0	377
	3	623	0	1412

分析源数据集得知，我们感兴趣的成绩处于及格边缘的学习者元组数据非常少，数据集存在类失衡的问题，必须对算法模型进行改善，以提高分类准确性。鉴于多类任务的类不平衡问题采用过抽样和欠抽样效果不明显，我们将成绩泛化成两类，即分数(grade)在0.5至0.7的学生者为成绩合格边缘学习者，其余的学习者为非边缘学习者，再对模型的输入数据进行过抽样处理。重新构建模型，得到两种模型在测试样本下的混淆矩阵如表3所示。

表3 两类任务在测试样本下的混淆矩阵

Tab.3 Confusion matrix of the binary task under the test sample

模型	类别	预测值	
		1	2
神经网络	1	31457	1075
	2	4707	27793
CART决策树	1	31212	1320
	2	4855	27645

计算得到神经网络模型预测的准确率为91.11%，预测值为1的类(成绩合格边缘学习者)的召回率为96.70%；采用CART决策树模型预测的准确率为90.50%，预测成绩处于合格边缘学习者所属类的召回率为95.94%。可见神经网络的预测性能更好，选择该模型用于学习者成绩预测更适合。

4 结论(Conclusion)

在线学习数据分析流程的研究对数据价值发掘意义匪浅。要发掘在线学习数据的潜在价值，满足在线学习数据分析需求，必须制定一个尽量完善的大数据分析方案，用来指导整个数据分析项目流程的执行。本文研究了数据分析前必须了解的分析对象相关信息，分析方案的详细制定过程及各个阶段的主要任务和方法，并且结合案例给出各个阶段的结果形式，对其他类似学习系统也具有推广与应用价值。然而，在本文的研究中还存在不足，本文为了简化预处理，缺失值采用了删除法，在具体的数据分析方案中需要研究采用删除法是否会影响数据集的数据结构，对此有待于今后更进一步理论和实验研究。

参考文献(References)

[1] 余鹏,李艳.大数据视域下高校数据治理方案研究[J].现代教育技术,2018,28(06):60-66.

[2] 罗达雄,叶俊民,郭霄宇,等.ARPDF:基于对话流的学习者成绩等级预测算法[J].小型微型计算机系统,2019,40(02):267-274.

[3] Jennifer M,Johanna F,Olaf K.Expectancy value interactions and academic achievement:Differential relationships with achievement measures[J].Contemporary Educational Psychology,2019,58:58-74.

[4] Wassan J T.Discovering Big Data Modelling for Educational World[J].Procedia-Social and Behavioral Sciences,2015,176:642-649.

[5] Ritanjali P,Ranjan S P,Dheeraj S.Online learning:Adoption,continuance,and learning outcome-A review of literature[J].International Journal of Information Management,2018,43:1-14.

[6] Prior D D,Mazanov J,Meacheam D,et al.Attitude,digital literacy and self-efficacy:Flow-on effects for online learning behavior[J].The Internet and Higher Education,2016,29:91-97.

[7] What is R?[EB/OL].https://cran.r-project.org/manuals.html.2019-04- 23.

[8] 高志鹏,牛琨,刘杰.面向大数据的分析技术[J].北京邮电大学学报,2015,38(03):1-12.

作者简介：

程 香(1982-)，女，硕士，高级工程师.研究领域：信息管理，数据分析研究.

程长征(1979-)，男，博士，教授.研究领域：计算固体力学.