doi: 10.3969/j.issn.1671-9247.2023.06.018

数据挖掘在高校试卷分析中的应用探究

许　青，王英明
（马鞍山学院，安徽 马鞍山 243100）

摘　要：以普通高校试卷题型和学生答题数据为研究对象，引入试卷分析中的数据挖掘思想，以期解决与实际应用密切相关的教学改进问题。将数据挖掘技术应用到试卷分析中，通过试卷中的相关数据更高效展现学生对不同知识点的掌握情况，并且预判试卷出题的合理程度，从而为教师进一步改进教学提供参考。

关键词：试卷分析；数据挖掘；评价
中图分类号：G642.0　　**文献标识码：**A　　**文章编号：**1671-9247(2023)06-0072-04

Exploring the Application of Data Mining in the Analysis of Test Papers in Colleges and Universities

XU Qing, WANG Yingming
（Ma'anshan College, Ma'anshan 243100, Anhui, China）

Abstract: Taking the question types of examination papers and students' answer data in general colleges and universities as the research object, the idea of data mining in examination paper analysis is introduced in order to solve the problem of teaching improvement which is closely related to the practical application. By applying data mining technology to test paper analysis, we can show students' mastery of different knowledge points more efficiently through the relevant data in the test paper, and predict the reasonable degree of questions in the test paper, so as to provide reference for teachers to further improve teaching.

Key words: test paper analysis; Data mining; evaluation

一、引言

普通高校教师在一门课程结束时会通过考试来评估学生成绩和检验教学效果，考试的试卷质量尤为重要。为了提高试卷的质量，一些教师会选择对往年的试卷进行分析，结合自身教学实际情况去设计较为合理的试卷。考试后对试卷进行分析还可以帮助教师了解教学效果，确定教学目标、提高教学质量[1]。

对于试卷分析的方法有很多，有通过学生及格率来计算客观题难度的；有采用平均分来计算主观题难度的；还有采用极端分组法来计算难度的[1]。方法有多种，但并没有一个统一的标准，更多的教师选择经验法，即通过总结以往的出卷经验来进行试卷分析。这种试卷分析方式往往只能发现卷面上存在的部分问题，不能全面反馈出试卷中可能存在的隐性问题。

数据挖掘技术（DM）能够通过大量数据，运用其算法挖掘出人们所需要的信息。在如今的信息社会，人们日常生活中随时都会产生大量的数据，如聊天记录、购物信息、网页浏览记录、小视频观看记录等，这些数据会以各种形式存储在网络上，方便商家和用户查看使用。这些数据最初可能会被忽略，但很快便有人关注，比如超市的购物清单会向商家展示一个人很多秘密，如个人喜好或者生活状态等，这些都可以通过数据分析被发掘出来。往往一些重要的信息都被隐藏在一堆数据中，而如何从大量的生活数据中寻找出对人类有用的信息，就成为了一个需要解决的问题，数据挖掘技术便很好地展示了这方面的优势。

当今社会，很多行业都采用数据挖掘技术来进行相应的数据研究。教育行业中，学者们也逐步开始思考如何利用数据挖掘技术来优化教育以及管理教学。本文通过对普通高校的某门课程无纸化试卷进行分析，利用数据挖掘技术，对试卷题型及学生答题情况进行分析，以便发现试卷中存在的潜在问题并加以改进，还可以帮助教师了解学生学习情况，以便教师进行有针对性的教学质量改进。

二、国内外的研究现状和发展趋势

（一）数据挖掘

在信息技术快速发展的当下，数据挖掘技术日益成熟起来，融合了多个领域的理论和技术（如数据库技术、机器学习、统计学、人工智能等），并且被广泛应用在多个领域。天文学家和加州理工学院合作开发了一个 SKICAT 系统，该系统能够帮助人们发现遥远的类星体，这也是数据挖掘技术在天文学上的首批成功应用之一。除此之外，商业中经常使用数据挖掘技术去分析客户的购买模式，“啤酒与尿布”就是一个典型的例子。沃尔玛超市工作人员通过分析购物清单，发现啤酒和尿布经常出现在同一个购物篮中，于是超市便通过改变啤酒与尿布的摆放位置，从而提高销售额。生物学家也在 DNA 的研究中使用了数据挖掘技术等。

在算法方面，美国的 IBM 公司早在 1996 年就研发出智能挖掘机，用于提供数据挖掘的解决方案。1997 年思维机器公司开发了 Darwin 套件，此外 SAS、Ora-

收稿日期：2023-03-02
基金项目：安徽省 2020 年度高校优秀人才支持计划项目：基于数据挖掘的高校试卷分析与应用（gxyq2020096）；安徽高校自然科学研究项目：基于大数据平台的舆情计算方法与技术研究（KJ2019A0916）
作者简介：许 青（1984—），女，安徽滁州人，马鞍山学院讲师，硕士。

cle 等公司也纷纷开发了相关产品。2005 年，怀卡托大学的 Weka 小组制作的 Weka 系统在当时也是得到了广泛认可的，该系统被视为数据挖掘历史上的一个里程碑，也是至今历史上最完备的数据挖掘工具之一。

国内直到 20 世纪末才开始对数据挖掘领域展开研究。检索中国知网有关数据挖掘方面的论文发现，2000 年仅有相关论文两百余篇，而 2004 年已上升到两千多篇。这些论文仅限于学术研究方面，相关的研究也主要都集中在对算法和应用方面的研究，研究的主体主要集中于以华为、阿里巴巴等为代表的金融、互联网等行业[2]。

（二）教育数据挖掘

当数据挖掘技术在商业、生物医疗、电信通讯等方面得到广泛应用后，人们开始尝试着将该技术与教育领域相融合，通过算法去挖掘教育数据中隐藏的有用信息。2000 年“教育数据挖掘（EDM）”这一概念出现。研究者成立了国际教育数据挖掘协会，该协会在 2008 年召开了第一届“教育数据挖掘国际会议”，并创办了一个在线学术期刊 *Journal of Educational Data Mining*（教育数据挖掘杂志）。

EDM 的出现立刻得到多方关注，美国教育部于 2012 年发布了《以数据挖掘和学习分析促进教与学》的报告，大篇幅介绍了如何通过数据挖掘相关知识来促进教育，以及就如何提高学生的学习效果进行了相关的理论和案例等多方面的分析和评估，不少高校也纷纷效仿并展开了相关研究计划。2014 年我国教育部教育信息化推进办公室颁布了《教育管理信息化建设与应用指南》，指南中明确指出各地要广泛应用教育管理系统，并对系统中存储的数据进行深度挖掘，从中发现学生学习和教师教学过程中存在的规律和问题，从而为推动教育教学改革提供重要的参考[3]。我国关于教育数据挖掘研究才起步不久，根据中国知网中的“教育数据挖掘”主题搜索结果，相关文献仅两百多篇，而且也是近几年才逐渐多起来的[4]。

教育数据挖掘是数据挖掘技术在教育领域中的应用，通过对研究领域和数据来源的分析可将教育数据挖掘分为四类：传统教学研究、网络教学研究、传统教务研究、信息化教务研究。教育数据挖掘的意义在于指导和改善学生学习情况，提高教师的教学质量。

美国的大学除了利用传统的考试对学生所学知识进行考核外，大部分教师都侧重对学生的学习行为进行评价，如实践动手能力、创新合作精神等。这些评价的结果有利于帮助学生提高学习水平。哈佛大学的研究者通过收集学生学习的相关信息，利用数据挖掘技术对其进行分析和研究，从而完成了对学生的相关评价，并为教师提供有用的反馈信息，从而帮助教师改进教学[5]。

国内对学生学习相关信息的收集，主要集中在对成绩信息的统计上，而关于学生对试卷中各知识点的得分情况的研究相对较少，这样就难以清晰呈现学生对相应课程中知识点掌握情况，也无法根据学生在试卷中的答题情况给相关教师提供反馈信息。

三、数据挖掘技术运用于试卷分析的实践

（一）K-Means 算法

本文采用的是数据挖掘聚类分析中的 K-Means 算法，该算法是一种无监督的学习，也就是给计算机一堆没有分类标记的数据，让计算机对数据进行分类、检测异常等。它是一种非层次的聚类算法，聚类中具体的类别数目需要在分析前就确定下来，整个过程是需要通过迭代的方式进行的。首先定一个初始的分类，然后通过不断迭代的方式把数据在不同类别中移动，直到最后达到一定的标准为止，整个过程具有很强的灵活性，不需要存储数据，因而不会出现多个互相嵌套的聚类结果，而且速度也很快[6]。

步骤一：先确定本次应用中的聚类数有多少，也就是 k 的值，这个数目是人为指定的。k 值的选择一直是个难题，有时面对大量数据根本不知道该如何分类以及该分多少类。常用的选择 k 值的方法有两种：（1）肘部法。这种方法比较适合于 k 值较小的时候，当选择的 k 值过大，数据中的样本对象就会被划分得很精细，同时对象的聚合程度也会增大，样本对象与它所属类别中心之间的距离（差异度）平方之和会渐渐变小，而这个平方和越小，则表示样本对象越接近它们的中心点，聚类效果越佳。在误差平方和变化的过程中，会出现一个拐点，整个曲线图像一个人的手肘，所以叫肘部法。（2）轮廓系数法。这种方法结合了聚类的凝聚度（对象和所属类别之间的相似度）和分离度（与其他类别做比较），它的目的是让样本内部距离最小化，外部距离最大化。先求出所有样本的轮廓系数，然后对这些轮廓系统求平均值就得到了平均轮廓系数，平均轮廓系数的取值范围为 [−1,1]，系数越大，聚类效果越好。当平均轮廓数接近 1 时，说明样本与所属类别之间联系密切，当轮廓数接近−1 时，则说明样本与所属类别间联系则疏远。

步骤二：根据上一步中找出的最优 k 值来初步设定每个类别的初始聚类中心。

步骤三：逐个计算对象到所属类别初始聚类中心之间的距离，将各个对象按照距离最近的原则归纳到相应类别中，同时计算出各类别的新聚类中心，也就是 means 的值。

步骤四：按照得到的新的聚类中心位置，重新计算出各对象距离新的聚类中心之间的距离，并且重新进行归类，更新类别聚类中心[7]。

步骤五：重复第四步，直到达到事先指定的迭代数为止。

（二）计算机基础课程试卷分析中的实践

1.确认对象

分析数据之前需要明确待解决的问题，了解所要研究的行业的数据和业务问题，如果缺少了这些就不能发挥数据挖掘的最终价值，从而也很难得到正确的结果。本文是以某普通高校大一学生大学计算机基础课程试卷为研究对象，对试卷卷面和学生的成绩进行整理和分析。

2.数据准备

数据准备是数据挖掘过程的核心阶段，可分为数据的选择、预处理和转换三个步骤。其中，数据的选择主要是实现对业务相关信息的搜索，并选择出可供数据挖掘使用的数据，以缩小处理对象的范围、提高数据挖掘的质量。数据的预处理步骤主要用以确定即将进行的数据挖掘类型以便为进一步的分析做准备。

数据的转换步骤则是建立特定的分析模型，将上面的信息转换成适用于相应挖掘算法的数据，该部分也是数据挖掘成功的关键。

本文的数据收集相对简单，计算机基础课程使用的是无纸化考试系统，可直接从系统中提取相应数据。因为系统中所涉及的专业学生较多，所以从中选择了 3 个具有代表性的数据：计算机专业的学生数据、非计算机专业工科的学生数据和非计算机专业文科的学生数据，下文分别以 a,b,c 代替进行区分。

因为无纸化考试系统中原始设置的题型过多，需要教师根据需求进行选择，本文选择将试卷题型分为单选题、中英文打字、Windows 操作、网络、Word 应用和 Excel 应用，其中各类题型的题量及单项总分，如表 1 所示。

表 1　各类题型的题量及单项总分表

考核项	单选题	中英文打字	Window操作	网络	Word应用	Excel应用	总计
题量	20	1	3	1	1	1	27
分值	30	10	15	5	20	20	100

根据分析需求，还需要对试卷中的各个题型的得分进行累加，统计出每个学生在各知识点上的得分情况，表 2 是随机选取的 8 位学生得分情况。

表 2　随机抽取的 8 名学生各题型的得分表

学生代号	单选	中英文打字	Windows	网络	Word	Excel
20407142	25.5	9.9	13	5	17.1	20
20407143	25.5	10	12.8	5	20	20
20407144	27	10	12	5	15.2	20
20407145	21	10	14.3	4.3	17.7	18.4
20407146	28.5	10	14	5	20	9.8
20407147	16.5	10	13.8	5	19.5	18.5
20407148	22.5	10	15	5	17.1	20
20407149	22.5	10	15	5	19.1	17.3

表 2 中的数据还要进一步进行转换，以适合进行 K-Means 算法的分析，根据试卷中各知识点的得分，给出每一项的分值比例，以小数形式表示，如表 3 所示。

表 3　小数形式表示的各项分值

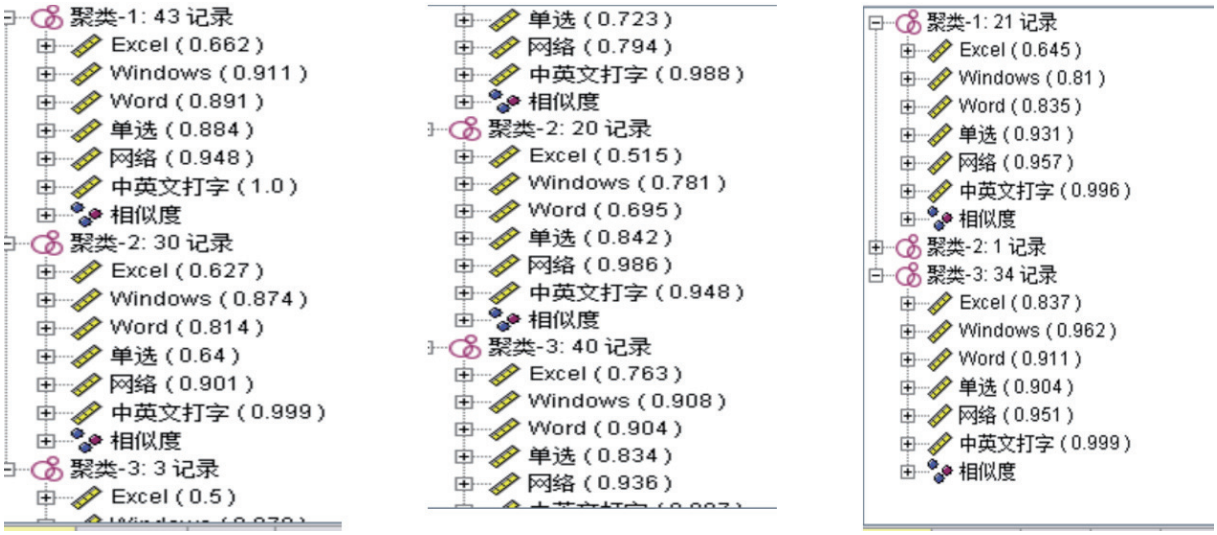
学生代号	单选	中英文打字	Windows	网络	Word	Excel
20407142	0.850	0.990	0.867	1	0.855	1
20407143	0.850	1	0.853	1	1	1
20407144	0.900	1	0.8	1	0.76	1
20407145	0.700	1	0.953	0.860	0.885	0.920
20407146	0.950	1	0.933	1	1	0.490
20407147	0.550	1	0.920	1	0.975	0.925
20407148	0.750	1	1	1	0.855	1
20407149	0.750	1	1	1	0.955	0.865
20407150	0.950	1	0.820	1	0.945	0.815

3.聚类分析过程

本文借助 Clementine 应用工具进行研究，该工具有可视化操作界面，对于熟悉 windows 操作的人来说，易学易用，比较适合于初学者实现对数据的挖掘。Clementine 中能够对已知对象数目的数据进行分类，然后收敛聚类，融合了数据挖掘的各种算法。本文根据其中的 K-Means 算法通过“划分”来实现聚类，这里选取的 k 值为 3，数据以范围类型进行导入，结果如图 1 所示。

4．聚类结果分析及指导分析

由聚类结果可以看出，a 类数据在聚类 1 中记录最多，其中 Excel 操作和选择题比例值相对较低；b 类数据在聚类 3 中记录为 40，Excel 操作比例值最低；c 类数据中聚类 1 和聚类 3 记录较多，同样 Excel 操作比例值



(a)　(b)　(c)

图 1　聚类结果

低；整体上看学生的中英文打字和网络部分的分值较高。

由此可以将学生在该门课程中各知识点的掌握情况反馈给相关的任课教师，以便后期的教学过程加以改进。如非计算机专业的学生 Excel 题分值较低，说明他们对于表格操作掌握情况不佳，教师可以在后期的教学中适当强化训练，以提高学生的动手操作能力。

四、结语

通过数据挖掘技术进行试卷分析，可以让教师了解试卷的合理性，从而有效促进教师的教学水平。教师结合试卷分析的结果对教学活动加以改进，有助于教师进一步提高教学质量。

参考文献：

[1]张秀娟. 试卷分析算法研究与应用 [D]. 北京：北方工业大学，2008.
[2]陈卓民. 数据挖掘技术在国内外的研究和发展现状 [J]. 青年文学家，2009(16)：122-123.
[3]雷晓锋，杨明. 教育数据挖掘的研究进展与趋势 [J]. 北京航空航天大学学报 (社会科学版)，2018，31(4)：108-114.
[4]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，2010，20(10)：21-25.
[5]常桐善. 数据挖掘技术在美国院校研究中的应用 [J]. 复旦教育论坛，2009(2)：72-79.
[6]张文彤，钟云飞，王清华. IBM SPSS 数据分析实战案例精粹 [M]. 2 版. 北京：清华大学出版社. 2020：5.
[7]张文彤，董伟. SPSS 统计分析高级教程 [M]. 3 版. 北京：高等教育出版社，2018：1.

(责任编辑　文双全)

（上接第 59 页）

（四）升级数控设备

数控设备的升级是工业 4.0 背景下数控实验教学不可或缺的一部分。开放式数控系统、多轴数控机床、高速加工中心、柔性加工辅助设备等是当前发展迅速的机床设备。高校应积极引入上述新的数控设备，构建全新的数控实验室，将传统的机床设备通过现代信息化技术链接在一起，加强开放式数控机床的二次开发管理和柔性化设备的适应性设计，在此基础上，逐渐引入数字孪生实训平台，解决传统数控实训成本高昂、场景固化、教学效果不佳等问题。

四、结语

数控技术日新月异，智能化、高速化、高精度化、并联驱动化、网络化、绿色化已成为数控机床发展的趋势和方向。传统的数控实验教学无法适应现代数控技术的发展，高校要更新课程内容、改进教学模式来满足技术发展的需求。教育要适应时代发展和社会需求，教师要不断的学习新的知识、更新教学理念，发展新的教学策略，并结合新的教学技术来提高教学质量与水平，着力培养学生实践能力、工程能力、动手能力以及创新能力。

参考文献：

[1]刘强. 数控机床发展历程及未来趋势 [J]. 中国机械工程，2021，32(7)：757-770.
[2]吴言政. 数控技术在智能制造中的应用现状及发展路径 [J]. 中阿科技论坛 (中英文)，2021(7)：35-37.
[3]李丽萍. “数控技术”课程教学改革探讨 [J]. 南方农机，2020(17)：146-148.
[4]杨梅生. 工业 4.0 背景下液压传动系统课程教学实践探索 [J]. 安徽工业大学学报 (社会科学版)，2022，39(2)：61-63.
[5]郭满荣. 数控铣床实践教学改革探索 [J]. 安徽工业大学学报 (社会科学版)，2013，30(2)：112-126.
[6]周凯，康剑梁. 智能制造背景下的数控实训教学研究与实践 [J]. 现代制造技术与装备，2018(10)：219-220.
[7]冯涛，邱昕洋，刘志杰，等. 基于创新能力培养的数控技术课程教学改革与实践 [J]. 教育教学论坛，2016(13)：64-65.
[8]谭霖. 基于信息化的数控技术教学改革初探 [J]. 教育现代化，2018(41)：18-19.
[9]任杰宇，李卫国. 数控加工虚拟仿真实验教学方法的改革与实践 [J]. 创新创业理论研究与实践，2021，4(17)：33-35.
[10]马岩，关宇. 谈校企合作在高校共享实验室平台建设中的积极作用 [J]. 知识文库，2017(9)：43-44.

(责任编辑　文双全)