基于 Apriori 数据挖掘算法的应用与实践

徐建军,张国华
(南京师范大学 泰州学院,江苏 泰州 225300)

摘 要:精准教育从诞生之初就受限于技术发展缓慢,随着信息技术的大力发展,教学管理系统,学生自主学习 APP 系统,基于微信学习平台的广泛应用,促使数学课程教育方面数据快速增长,使得学生的学习的行为,过程,状态,练习结果,成绩等成为可以被信息技术自动抓取的数据存在,这样使得获取精准教学的测量数据更为便捷和有效。首先分析了数学课程教学活动中存在的问题,然后对信息化教学模型构建进行了分析。以大数据中 Apriori 算法为主要思想,设计了基于学生数学学习效果提示和老师教学效果预测功能的数据挖掘系统,实现了对可能学习效果不理想的学生和可能教学方向不精准的老师的及早提示,同时可避免学生教师过度重复已掌握很好的知识,精准定位每位学生薄弱环节并加以改进,既可减轻师生负担,又可提高教学效果。

关键词:数学课程;精准教学;数据挖掘;Apriori 算法
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2020)04-0206-05
doi:10.3969/j.issn.1673-629X.2020.04.039

Application and Practice of Data Mining Algorithms Based on Apriori

XU Jian-jun, ZHANG Guo-hua
(Nanjing Normal University Taizhou College, Taizhou 225300, China)

Abstract: Precision education has been limited by the slow development of technology since its inception. With the vigorous development of information technology, teaching management system, students' self-learning APP system, and the wide application of Wechat learning platform, the data of mathematics curriculum education has increased rapidly, which makes students' learning behavior, process, state, practice results and achievements possible to be grasped automatically by information technology. The data obtained exist, which makes it more convenient and effective to acquire the measurement data of precise teaching. Firstly, we analyze the problems existing in the teaching activities of mathematics courses, and then analyze the construction of information-based teaching model. Taking Apriori algorithm in big data as the main idea, a data mining system based on the function of prompting students' learning effects and predicting teachers' teaching effects is designed, which realizes the early prompting for students with unsatisfactory learning effects and teachers with inaccurate teaching directions, and at the same time, can avoid students and teachers over-duplication of superb grasp of knowledge. It can accurately locate each student's weak links and improve them, which can not only reduce the burden of teachers and students, but also improve the teaching effect.

Key words: mathematics course; precision teaching; data mining; Apriori algorithms

0 引 言

随着人工智能、大数据、云计算、数据挖掘等技术的深入发展,信息技术已成为各行各业不可或缺的工具。而教学是未来人才培养的基础学段,面对这样一个充满个性发展的时代,主动开展基于信息技术的精准教学,是教育发展的必然趋势。

精准教学[1]是学者 Lindsley 在 20 世纪 60 年代在学者 Skinner 的行为学习理论的基础上提出的一种新的教学方法。精准教学诞生之初就是用于靶向教育,其目的是通过设计测量教学过程来获取有关数据,以追踪学生的学习表现并为教学提供数据决策支持。其衡量标准主要是流畅度。流畅度是由学生对应掌握知识和技能的“准确程度”和“熟练速度”两方面组成。而传统的精准教学的监测办法是通过人工绘制测量表来分析,过于繁琐,且缺乏大数据和人工智能为支撑,结果往往不够精准。

收稿日期:2019-05-27 **修回日期**:2019-09-27 **网络出版时间**:2019-12-18
基金项目:教育部 Google2014 年产学合作专业综合改革项目(PO640068)
作者简介:徐建军(1981-),男,讲师,硕士,研究方向为计算机网络及大数据。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191218.1113.060.html

文中采用大数据与人工智能技术,基于数据挖掘理论,采用了先进的挖掘算法[2],以多年教学过程中学生数学练习及部分模拟数据作为数据源,智能分析出学生不同学段知识点之间的关联性,及时给学生学习预警提示,给教师教学效果预测提醒,使得学生学习教师教授新知识点的时候,能够得到有价值的推送信息。

例如,学生会清楚地看到前期哪些知识点与该知识点的关联度高,影响度又是如何的,同时,教师在讲解该知识点之前,会出一个统计报表,预测常规方法教学,班级学生掌握情况,能够精确定位班级每名学生学习该知识点造成影响的前期薄弱环节,及时高效解决。

1　国内外精准教学研究现状及应用困境

1.1　国内外精准教学研究现状

经查阅相关国外文献,国外的精准教学主要研究点集中在通过具体的教学实验及案例来评估教学效果。

例如 Downer 和 Griffin 两位学者的研究结果表明,精准教学对提高学生的阅读能力具备明显优势[3];而 Gallagher 和 Stromgren 两位学者的研究结果则表明精准教学对帮助数学学习困难的学生具备优势[4-5]。

在国内,精准教学研究处于起步阶段,笔者在中国知网中尝试以“精准教学”为关键词进行检索(数据截至 2019 年 5 月 27 日),结果为 44 条,相关研究成果大都集中在近两年。

比较典型的成果[6-7]有,梁美凤发表了《“精准教学”探析》;祝智庭教授发表了《信息技术支持的高效知识教学:激发精准教学的活力》;王永雄、丁德瑞等学者发表了《基于创新实践能力培养的精准分层教学》等。

由此可知,目前国内的精准教学研究总量偏少,精准教学的研究成果不多,大数据、云计算、人工智能的分析方法应用较少。

1.2　精准教学应用困境

精准教学虽然是一种有效的教学方法,但在实际教学过程中却很难应用好,分析原因主要有两点:

(1)精准教学是通过设计测量表来获取学生学习数学的行为结果,并依据测量结果进行薄弱环境的强化训练,从而提高学习质量。该方式缺乏对每名学生学习过程的有效监控,往往会忽略学生的个性化发展。

(2)精确教学缺乏先进的技术支持。精准教学主要以测量,收集学生学习表现数据为数据基础,以统计学相关知识为技术基础。而数学课程教师往往采用传统的人工管理数据方式进行分析,大多缺乏对于先进统计学,数据分析决策工具的应用,因此数据采集效率,可视化及精准化程度均不高。

2　大数据及数据挖掘技术对精准教学的影响

谁掌握数据,谁就掌握主动权,大数据的兴起正引领社会发展新变化。大数据技术必将给精准教学提供更多更先进的理论与技术支撑,促进其大力发展。

2.1　大数据技术使得数学课程中采用精准教学更为可行

随着信息技术的大力发展,教学管理系统,学生自主学习 APP 系统,基于微信学习平台的广泛应用,促使数学课程教育方面数据快速增长,使得学生的学习的行为,过程,状态,练习结果,成绩等成为可以被信息技术自动抓取的数据存在,这样使得获取精准教学的测量数据更为便捷和有效。

2.2　精准教学可以使用先进的数据挖掘技术

数据挖掘技术(data mining),是指从海量数据中自动发现隐藏于其中有某种关联的信息过程。在获取了精准教学的基础大数据之后,就可以采用先进的数据挖掘技术来分析每个学生的薄弱点,精准定位学生知识学习的薄弱环节,给出相应的学习方案。同时大数据支持多并发及海量数据的能力,教师就可以使精准教学规模化地应用到每名学生身上,兼顾学生的个性化发展。

2.3　大数据技术的广泛使用可以使得精准教学更加开放智能

大数据[7]一般满足容量大(Volume),多样性(Variety),速度快(Velocity),有价值(Value),真实性(Veracity)五个特性。在教育领域,大数据平台是服务于教育教学工作的综合性的信息平台,师生、家长等角色均是大数据的生产者和应用者。因此,随着大数据技术、数据挖掘技术在数学课程精准教学中的应用,学生、教师、家长等角色均可参与到精准教学过程中,可以为学生量身定制教学方案,家长随时可以掌握自己孩子的学习情况,教师可以掌握班级整体情况,教学管理者可以根据大数据把控教学改革方向等。同时大数据具备自我学习能力,随着数据量的不断增加和数据挖掘算法的不断改进,其数据分析处理能力将更加智能精准。

3　基于数据挖掘支持下的精准教学模式设计

在现有的教学环境下,数学课程教师更加倾向于使用成熟的教学模式,精准教学往往被作为是评估教学效果的策略或方法,老师们也似乎不太愿意花费多余精力来研究如何将精准教学融入现有教学中,而是照搬成熟的教学模式来使用,所以精准教学并没有被

广泛采纳。此外,在信息化引领的教育变革潮流中,精准教学也因为信息技术支撑的缺失而受到广大教育工作者的冷落,缺乏活力。大数据可以突破传统教学限制,大力推动数学课程教师在思想和行为上接纳并认同精准教学,利用大数据来构建可供参考精准教学的模式,可以推动精准教学发展、促进精准教学应用。为此,文中将从精准教学目标、程序化教学过程、基于数据挖掘的评价与预测[8]三个方面,来构建数学课程大数据的精准教学模式框架,具体如图 1 所示。

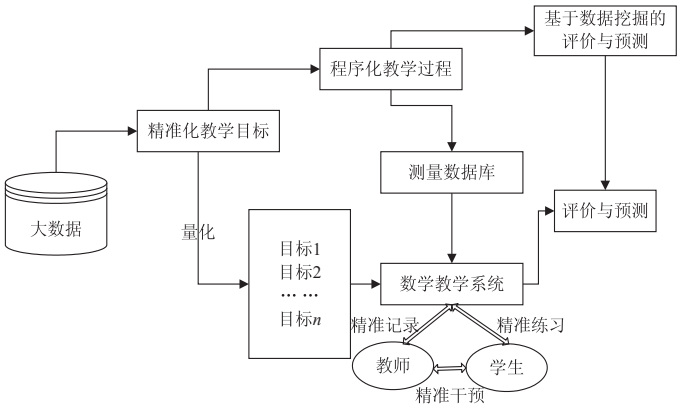

图 1 数学课程大数据精准教学模式框架

3.1 精准教学系统总体架构

精准教学信息化平台采用标准三层架构设计,其中表示层负责数据采集、修改、删除,智能设定数据挖掘算法的可信度与置信度[9]最小值,对于学生可自动提示后续可能学不好的知识点,对于老师则可自动预测该班的整体教学效果,并分析出关联的知识点。业务逻辑层主要实现关联挖掘算法,快速对大数据进行挖掘,并反馈结果。数据层主要采用主流的大数据框架 Hadoop[10]以及强大的 MapReduce 来实现存储原始数据及并行处理,建立挖掘数据库,形成关联规则数据库,其标准三层架构如图 2 所示。

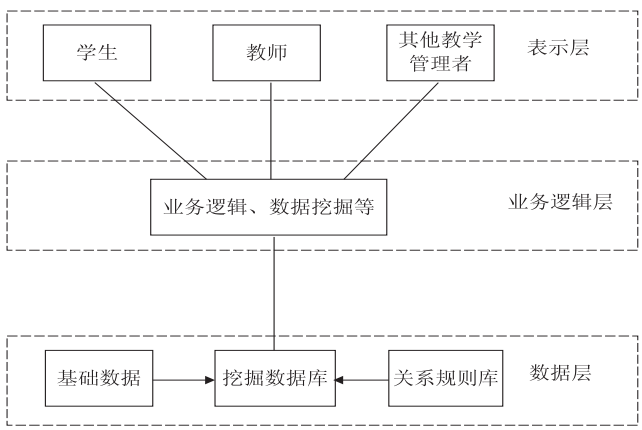

图 2 精准教学系统整体架构

3.2 精准教学目标

明确教学目标是整个教学活动的关键,是检验教学是否成功的依据。精准教学的第一步是必须精准化教学目标。例如在《数学课程课程标准》的第二学段中关于“数的运算”中要求“探索并了解运算律,会应用运算律进行一些简便运算”,这里的“会应用”是一个相对比较笼统的词语。

在精准教学体系中,可以建立更明确更精准化的教学目标,对每个教学目标都可转化成可以被信息系统抓取的可细化和量化数据。把传统的“会应用”、“熟练掌握”等这些较为笼统的教学目标,经过细化、分解、量化之后,建立起既包含准确掌握知识技能的目标,又包含运用知识技能速度的目标,达到精准教学的“流畅度”要求的指标。

3.3 程序化教学过程

精准教学诞生于 Skinner 的程序化教学过程,程序化是精准教学的关键。文中基于大数据程序化教学过程,具体包括:(1)建立数学课程大数据平台,利用精准化、智能化推荐技术,根据学生的不同特点,配置对应教学资源,融入个性化的教学;(2)改进传统的教学过程,收集每位学生的学习过程数据,为下一步数据挖掘,决策和干预做好准备;(3)实施精准干预,根据大数据平台反馈的结果,结合精准化的教学目标自动判别学生是否达到要求,如果有问题,则需要干预,会自

动回溯,精准定位到薄弱点及关联点,实现精准干预。

3.4　Apriori 算法

3.4.1　基本定义

定义 1(可信度):设 $I=\{i_1,i_2,\cdots,i_n\}$ 是含有 N 项的集合,交易 T 是项的集合, $T\subseteq I$, D 为 T 集合,设 A 是其中项集,事务 T 包含 X ,仅有 $X\subseteq T$,即 X 包含于 T , $X\Rightarrow Y$ 是关联规则表达式,且 $X\subset I$, $Y\subset I$, $X\cap Y=\varnothing$。

$$\text{Support}(X\Rightarrow Y)=P(X\cup Y)=\{T:X\cup Y\Rightarrow T,\ T\in D\}/D\times 100\%=a \tag{1}$$

定义 2(置信度):关联规则 $X\Rightarrow Y$ 的可信度是指包含 X 与 Y 的交易数与仅包含 X 的交易数的比值,表示为:

$$\text{Confidence}(X\Rightarrow Y)=P(X\mid Y)=\frac{\text{Support}(X\cup Y)}{\text{Support}(X)}\times 100\% \tag{2}$$

3.4.2　Apriori 算法

Apriori 算法[11]是一种依据频繁项集作为先验知识[12],采用逐层扫描的迭代方案,即第 k 项集用来检索第(k +1)项集。具体步骤如下:首先通过检索事务(或交易)的记录,扫描出所有频繁的第 1 项集,并将找到的集合记为 L_1 ,依次迭代寻找出 L_2 , L_3 …如此循环下去,直到找不到任何频繁的第 k 项集。其次利用所有的频繁集中提取的强规则,即能为用户决策提供支撑的关联规则。其相关为代码如下:

```
输入:D 是事务数据库;min_sup 是最小支持度技术阈值
输出:L 是事务数据库中的频繁项集
方法:
L1 = find_frequent_1-itemsets( D );
//首先找出所有频繁 1 项集
For( k = 2;Lk - 1! = null;k ++ )
{
Ck = apriori_gen(Lk - 1)
//循环产生候选同时剪枝
For each 事务 t in D
{
//通过循环对 D 扫描进行计数
Ct = subset( Ck,t)
//调用得到 t 的子集
For each 候选 c 属于 Ct
c. count++;
}
Lk = { c 属于 Ck | c. count>=min_sup}
}
Return L=得到所有的频繁集;
```

3.4.3　前期数据准备

以某校 5(1)班 60 名学生为研究对象,选取数学四年级以上学期课程成绩数据为例,通过相应的数据挖掘算法分析,找到知识点的内在关联,为学生的学习,老师的教学提供数据决策支撑。为方便表述,将初步认识平面图形特点成绩表示为 PMTX,认识立体图形的体积表示为 LTTX,表 1 给出了以学号为主键成绩汇总。

表 1　学生数学成绩汇总(分)

学号	单元与知识点				
	PMTX	LTTX	…	…	…
00000001	68	79	73	79	…
00000002	60	49	69	45	…
00000003	76	75	69	85	…
…	…	…	…	…	…
00000060	53	72	75	64	…

3.4.4　数据挖掘筛选、转换

该班 60 名学生中有 1 名学生转学,1 名出国,后续知识点的单元无关联,因此从挖掘数据库中排除,得出有效成绩元组 964 条(平均每学期选取了 4 个知识点)。为方便数据挖掘,需要对数据进行格式化处理,成绩与等级关系如表 2 所示。

表 2　成绩与等级关系

成绩	等级
90-100	A
80-89	B
60-79	C
59 以下	D

转换成挖掘数据如下:

根据表 3,对照 Apriori 算法设定最小支持度为 0.2,班级总人数为 58 名,那么最小支持人数必须达到 52 * 0.2 即 11.6 名,如果这个知识单元的所在等级的人数不足 11.6 名,则需排除在数据挖掘库中,因此设定最小支持度为 0.2 时选取的数据如表 3 所示。

表 3　学生成绩汇总(人次)

等级	PMTX	LTTX	…	…	…
A	16	13	12		…
B	17		18	14	…
C	15	21			…
D				18	…

3.5　基于数据挖掘的评价与预测分析

(1)对于学生的评价和预测。

通过部分真实及模拟数据,设置数据挖掘时支持度≥0.8,置信度≥0.7,利用改进数据挖掘 Apriori 算法,得到以下 53 条关联规则,具体如表 4 所示。

学生登录系统后,自动获得一些推送信息[13]。例如,该生第一学段的“初步认识平面图形特点”该知识

点考了 83 分,位于‘82.5 ~ 87.5’区间,该大数据平台将预测到按传统方法学习第二学段的“认识立体图形的表面积”将 75% 的可能取得的分值位于‘81.0 ~ 86.0’区间,效果不理想,系统会自动回溯,精准定位影响本知识点学习的知识点,实现精准干预,学生能够及时轻松弥补欠缺知识点,减少后续影响[14]。

表 4 数据挖掘结果关联规则表

序号	关联规则	置信度
1	第一学段:小数分数的认识 = ‘81.5-86.5’ => 第二学段:小数分数及百分数意义 = ‘80.5-85.5’	0.73
2	第二学段:用字母表示数 = ‘80.5-85.5’ => 第二学段:利用方程解决实际问题 = ‘80.5-84.5’	0.72
…	…	…
53	第一学段:初步认识平面图形特点 = ‘82.5-87.5’ => 第二学段:认识立体图形的表面积 = ‘81.0-86.0’	0.75

(2)对于教师教学的预测和改进。

老师登录系统后,系统会给出班级学生信息,给出对本知识点造成影响的已学知识点信息[15],同时也会罗列出可能将对哪些后续知识点造成影响,并给出本知识点对于数学课程后续知识点影响程度及关联程度。如果预测班级整体学生 80 分以下比例(置信度≥0.75 统计)占比大于等于 20%,则为严重,大于等于 15% 为一般,大于等于 5% 则为轻微,如果预测影响后续知识点数目(置信度≥0.75 统计)大于等于 3 个为严重,大于等于 2 个为一般,大于等于 1 个为轻微,老师将会看到如下类似统计表,如表 5 所示。

表 5 知识点“认识立体图形的表面积”教学效果预测

学 号	已学知识点(分值,置信度)	未学知识点(分值<80)(置信度)
00000001	无	无
00000002	初步认识平面图形特点(83,0.77)	认识立体图形的体积(0.76)
…	…	…
80 分以下比例预测	10/48	影响后续单元数 3
影响等级	√严重 □一般	□轻微
关联等级	√严重 □一般	□轻微

通过对学生和教师两方面的预测和评价,注重给学生精准引导和补习前序单元知识点的缺失,又能提醒老师主动调整教学方法来实现精准教学,双方共同配合,教学效果能得到大幅提升。

4 结束语

精准教学是一种先进的教学模式,可以大大减轻学生的学习负担,提升学习效率,避免教师过度重复教学[2],在明确教学方向和内容方面具备优势,但是精准教学如果缺乏了大数据技术,人工智能,数据挖掘技术的支持,就很难发挥出效果。文中设计并模拟了精准教学的数据挖掘平台,可以有效发挥精准教学的优势。鉴于部分数据采用模拟数据,后续工作将继续展开研究,进一步收集数据,充分发挥精准教学优势。

参考文献:

[1] STRØMGREN B, BERG-MORTENSEN C, TANGEN L. The use of precision teaching to teach basic math facts[J]. European Journal of Behavior Analysis, 2014, 15(2): 225-240.

[2] BACA R, KRATKY M. T3Dewey: on the efficient path labeling scheme holistic approach [M]//Database systems for advanced application. [s. l.]: Springer, 2009: 6-20.

[3] 陈崇成,林剑峰,吴小竹,等. 基于 NoSQL 的海量空间数据云存储与服务方法[J]. 地球信息科学学报,2013,15(2):166-174.

[4] 郑怡文,陈红星,白云晖. 基于大数据在课堂教学中对学生精准关注的实验研究[J]. 现代教育科学,2016(2):54-57.

[5] 钟绍春. 教育云、智慧校园和网络学习空间的界定与关系研究[J]. 中国教育信息化,2014(3):3-8.

[6] 梁 盾. 数据挖掘算法与应用[M]. 北京:北京大学出版社,2007:35-42.

[7] 孙爱婷,李斯远. 大数据时代的教育方式改革[J]. 中国管理信息化,2017,20(5):244-246.

[8] 申德荣,于 戈,王习特,等. 支持大数据管理的 NoSQL 系统研究综述[J]. 软件学报,2013,24(8):1786-1803.

[9] 胡水星. 大数据及其关键技术的教育应用实证分析[J]. 远程教育杂志,2015,33(5):46-53.

[10] 段金菊,余胜泉. 学习科学视域下的 e-Learning 深度学习研究[J]. 远程教育杂志,2013(4):43-51.

[11] 张国华. 数据挖掘在独立学院课程预警中的应用与实践[J]. 现代电子技术,2016,39(17):136-139.

[12] 林 建,董亚波,朱淼良. 基于 J2EE 的数据访问技术设计模式研究[J]. 计算机工程与应用,2004,40(14):129-131.

[13] VOGELS W. Eventually consistent[J]. Queue, 2008, 6(6): 14-19.

[14] ELMASRI R, NAVATHE S B. Fundamentals of database systems[M]. 孙 瑜, 译. 4th ed. 北京:机械工业出版社,2009.

[15] GILBERT S, LYNCH N. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services[J]. ACM SIGACT News, 2002, 33(2): 51-59.