Advances in Education 教育进展, 2023, 13(3), 1255-1262
Published Online March 2023 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2023.133198

Hans汉斯

数据挖掘方法在大学通识课建设中的实践研究

——以《计算机网络技术与应用》课程为例

王行建，刘　欣*

东北林业大学信息与计算机工程学院，黑龙江 哈尔滨

收稿日期：2023年2月17日；录用日期：2023年3月17日；发布日期：2023年3月27日

摘　要

为了提升通识课教育的教学方法和改进通识课培养方式，本文以某高校《计算机网络技术与应用》课程为例，在大量教学数据的基础上，采用关联规则的数据挖掘方法，对课堂测试成绩、作业完成度、任务点完成度、期末成绩等内容进行了关联性挖掘分析，获得了五条关联规则，发现了新的教学规律，文章还利用决策树算法，考虑不同学科、不同课程基础、不同人才培养方案设定等条件下，进行课程成绩预测，从而为学业预警提供了有力的支撑。

关键词

关联规则，决策树，学业预警，通识课

The Practical Research of Data Mining Method in the Construction of Liberal Arts Education in Universities

—Taking “Computer Network Technology and Application” as an Example

Xingjian Wang, Xin Liu*

College of Information and Computer Engineering, Northeast Forestry University, Harbin Heilongjiang

Received: Feb. 17th, 2023; accepted: Mar. 17th, 2023; published: Mar. 27th, 2023

Abstract

In order to improve the teaching and training of Liberal Arts Education courses, this paper takes

*通讯作者。

文章引用: 王行建, 刘欣. 数据挖掘方法在大学通识课建设中的实践研究[J]. 教育进展, 2023, 13(3): 1255-1262.
DOI: 10.12677/ae.2023.133198

the case of "Computer Network Technology and Application". Based on a large amount of teaching data, using the data mining method of association rules, it conducts association mining analysis on classroom test scores, homework completion, task point completion, and final grade, and obtains five association rules as well as discovers new teaching rules. The article also uses decision tree algorithms to predict course performance under the condition in different disciplines, different course foundations, and different talent cultivation schemes, providing strong support for academic warning.

Keywords

Association Rules, Decision Tree, School Precaution, Liberal Arts Education

1. 引言

通识教育是针对大学教育日益专业化、职业化而采取的一种补充和平衡措施[1]。通识教育可以充分体现学科之间的交叉融合、专业知识之间的联系互补。当前高校的发展建设，遵循学科齐全，多学科融合发展，本科阶段通识课设置在 200 多门左右兼顾自然科学类、人文社会科学类、艺术类等，打通了学院、学科的专业限制，研究生阶段也普遍设置了通识课程。但随着跨专业选课学习，不同年级，不同学科门类的同学，对于同一门通识课程的学习效果表现出较大的差别。

自疫情出现以来，我国高校响应教育部“停课不停学”的号召，开展了大规模的线上教学[2]。无论是单纯的线上教学，还是线上、线下混合式教学，都极大利用了网课平台，不但适应条件，改革了教学方式，更重要的是保存下了大量有用的数据信息，可进一步利用数据分析的方式，挖掘出数据之间的耦合关系，从而为教学大纲修改定制、针对性教学、灵活高效的考核，以及教师教学自查整改都提供了科学依据，奠定了改革基础。

2. 研究设计

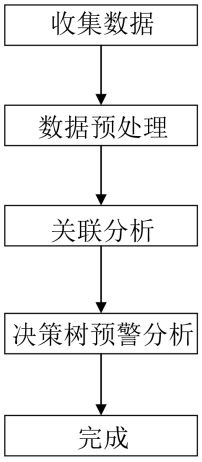

Figure 1. Diagram of analysis process
图 1. 分析流程图图

研究的目的在于利用现有的资源和方法，通过实例化的分析，找到隐藏在数据之间的关联性质和发展演化状态，从而为分析当前教学大纲指导下的教学改革提供参考和借鉴，同时为人才培养方案的制定、学业预警等提供有利的支撑。

本文以某高校“计算机网络与应用”通识课程为例，以课程授课过程中线上线下的课程教学数据为基础，利用关联规则与决策树算法的融合，对课程授课过程中各个环节的学习情况对总体成绩的影响进行分析，并且利用关联规则获取的强规则，作为决策树算法的新的分类属性，从而进一步研究学科分类因素对于成绩的影响，进而进行成绩的预测和学业预警。如图 1 所示。

3. 数据来源与预处理

3.1. 数据来源

除了一些众所周知的英文缩写，如 IP、CPU、FDA，所有的英文缩写在文中第一次出现时都应该给出其全称。文章标题中尽量避免使用生僻的英文缩写。

本文的研究对象是选修某高校通识教育选修课程“计算机网络技术与应用”的本科生，通过对 2016~2021 年选修本课程的学生学习过程相关数据进行了收集整理。

由于涉及的教学数据复杂而庞大，并不适合全部选取，针对线上线下相结合的教学方式，选取了来自不同专业的选修本课程的 261 人的教学相关数据作为研究目标。

通过收集整理相关数据形成了学生视频观看数据集、课堂测试数据集、作业完成度数据集、期末成绩数据集。

3.2. 数据预处理

学生观看的网课视频按照学习平台的记录方式分为时长与重复观看次数、作业的数据可以利用作业成绩标记。

数据清洗：对中途改选、退选等情况，可能出现作业提交缺失、未参加考试、视频仅观看一部分等情况，则删除此部分学生的记录。

数据离散化：根据课程编制的“考核标准”，将视频观看、作业成绩、课堂测试成绩、期末成绩分别利用不同离散值形式进行标定。

课堂成绩、期末成绩根据五级制通用形式，标记为“A1~A5”、“D1~D5”，任务点和作业并不做差异性区分，以完成数量进行标记，将任务点的完成度和作业的完成度，按完成的数量分别分为“R1~R5”、“B1~B5”五个等级，如下表 1 所示。

Table 1. Table of attribute discretization
表 1. 属性离散化对应表

属性	值	代码	属性	值	代码
课堂测试成绩	>90 分	A1	作业完成度	81%~100%	B1
	80~89 分	A2		61%~80%	B2
	70~79 分	A3		41%~60%	B3
	60~69 分	A4		31%~40%	B4
	<60 分	A5		<30%	B5
任务点完成度	26~30 个	R1	期末成绩	>90 分	D1
	21~25 个	R2		80~89 分	D2

Continued

16~20 个	R3	70~79 分	D3
11~15 个	R4	60~69 分	D4
<10 个	R5	<60 分	D5

4. 成绩关联规则相关性分析

4.1. 关联规则算法

关联规则是一种能够发现海量数据之间的隐含关系的数据挖掘技术，它是在上个世纪九十年代中期提出的，并一直活跃至今。

关联规则被广泛关注和使用，例如经典的啤酒和纸尿裤的案例，同时在教育行业分析中也常常见到利用关联规则获得教育教学过程中的隐含关系。

关联规则通常以关联数据和交易数据为载体，查找事物或者对象之间存在的频繁模式和关联规则[3]。

关联规则可以揭示项目或者对象之间内在联系，这些联系的强弱[4]，通常需要关键规则的三个指标来进行衡量：支持度(*Support*)、置信度(*Confidence*)和提升度(*Lift*) [5]。

项集就是一个或者 N 个项的集合，例如集合{100, 98, 70, 59}就是一个由 4 个项组成的集合。所谓的频繁项集为某一项集出现的频率达到设置的最小支持阈值。

通常把集合中项集 A 出现的概率认为是项集 A 的支持度；项集 A 和项集 B 的支持度可以用两者同时出现的概率表示：

$$\begin{aligned} &Support\left(A\right)=P\left(A\right) \\ &Support\left(A\Rightarrow B\right)=P\left(A\cap B\right) \end{aligned} \tag{1}$$

置信度为在项集 A 存在的前提下，项集 B 同时存在的概率：

$$Confidence\left(A\Rightarrow B\right)=P\left(B\middle|A\right) \tag{2}$$

通过设定最小支持度和最小置信度来进行衡量和筛选，即为高于最小支持度的阈值为重要的，最小置信度阈值为最低要求。

4.2. Apriori 算法

Apriori 算法和改进的 Apriori 算法是当前关联规则中使用最为广泛的方法之一，同时 Apriori 算法也是最先用于数据集挖掘的算法[6] [7]。

它的基本思想就是首先生成包含高于所设置的最小支持度的所有项集，即频繁项集，然后通过扫描事务数据库，反复遍历项集，利用迭代的方式，产生要得到的频繁项集。

算法得步骤如下图 2 所示。

第一步，设置最小支持度和最小置信度。

第二步，扫描事务集 D，如果满足产生候选集的条件，则产生候选项集，再选出满足最小置信度和支持度的频繁项集。

第三步，通过迭代的方式，反复的遍历候选集，更新频繁项集中的支持度。

第四步，生成关联规则。

4.3. 关联结果分析

利用筛选的学习课程原始数据，通过 Apriori 算法挖掘出“计算机网络技术与应用”课程的课堂测试

成绩、作业完成度、任务点完成度以及期末成绩之间的关系，得到如下表 2 的规则。

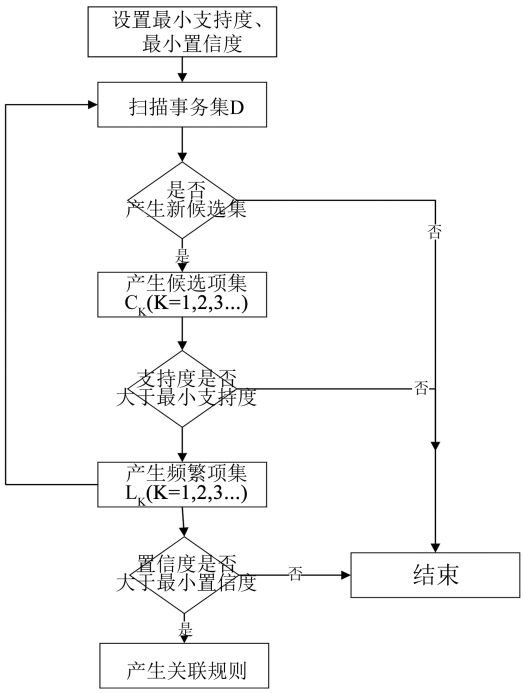

Figure 2. Diagram of the Apriori algorithm flow
图 2. Apriori 算法流程图

Table 2. Table of association rule result
表 2. 关联规则结果表

序号	前项	后项	支持度	置信度
规则 1	A3、B3、R4	D4	0.416	0.817
规则 2	A4、B3、R3	D5	0.396	0.758
规则 3	A2、B2、R2	D1	0.424	0.835
规则 4	A4、B2、R2	D3	0.415	0.754
规则 5	A4、B4、R4	D5	0.379	0.897

通过实验的结果得到了 5 条规则：

规则 1：课堂测试 = “A3”，作业完成度 = “B3”，任务点完成度 = “R4”，期末考试成绩 = “D4”，置信度为 81.7%。

规则 2：课堂测试 = “A4”，作业完成度 = “B3”，任务点完成度 = “R3”，期末考试成绩 = “D5”，置信度为 75.8%。

规则 3：课堂测试 = “A2”，作业完成度 = “B2”，任务点完成度 = “R2”，期末考试成绩 = “D1”，置信度为 83.5%。

规则 4：课堂测试 = “A4”，作业完成度 = “B2”，任务点完成度 = “R2”，期末考试成绩 = “D3”，置信度为 75.4%。

规则 5：课堂测试 = “A4”，作业完成度 = “B4”，任务点完成度 = “R4”，期末考试成绩 = “D5”，置信度为 81.7%。

根据上面的五条规则可知，当任务点完成度和作业完成情况达到一定时长，单元测试成绩较好，则期末考试的成绩高；当任务点学习和作业完成都一般的情况等下，单元测试成绩即便刚刚及格，期末成绩往往不及格，在任务点和作业情况达到中等偏下的情况下，单元测试的情况对期末考试成绩的影响是最大的。

从现实情况中看，在线学习和任务点的完成情况，可能与实际效果之间存在一定误差，这些误差往往是主观行为造成的，例如，存在刷时长，听而不闻，机械性的完成任务点的学习等等情况，这就使得完成情况与知识掌握情况不相符。课堂测试反应的情况相对更加客观一些，反映了对一些知识点的掌握情况，但是课堂测试的题量往往有限，覆盖面不大，难度不高，所有又不能完全反应对该段时间知识点的学习情况。

5. 决策树成绩预警预测

自从“学业预警”[8]制度实施以来，实现了对于学生学习情况的动态化管理，对于帮助学生完成学业，提升学校的教学质量和人才培养质量都有较大的帮助，但是目前“学业预警”通常是对于学生整体学习成绩的动态化管理，而对单科成绩的动态化管理却显不足。

通识课在人才培养体系中所占有的比重较大，尤其是学分绩点大的课程其成绩直接会影响到学业绩点情况，还会对学生心里和情绪产生一定的影响，可见，对于这样的课程成绩的动态关注与管理，对学生学业整体的影响是比较大的，利用成绩预警预测的方式来实现这一动态化的管理不失为一种好的办法。

由于通识课的特殊性质，同一门课程选修的学生可能来自于不同学科、不同专业，甚至工、理、文、管、农、林等学科都有涉猎，学生往往又来自不同年级，即便是同一专业，不同年级的学生培养方案也存在不同，这就注定学生基础知识储备不同；学生选课的目的也各不相同，有对课程内容知识感兴趣的、有为了满足学分要求的、有为了提升绩点的，这些隐藏的因素往往对成绩的影响巨大，利用决策树的分类方法提供了一种综合考虑多因素影响下的成绩预测方法。

5.1. C4.5 决策树算法

Apriori 算法虽然给出了较好的结果，但是它的迭代过程中需要反复的扫描数据库，而教学数据繁冗庞大，势必会造成算法效率低下、应变性差的特点。针对本研究，C4.5 决策树算法[9]在很大程度上与 Apriori 算法形成互补，C4.5 算法数据结构简单可以处理离散型和连续型数据，便于理解和掌握，对于处理效率高，结果简单明了、清晰易懂。

5.2. C4.5 算法计算方法

根据信息增益公式：

$$\begin{aligned} &Support\left(A\right)=P\left(A\right)\\ &Support\left(A\Rightarrow B\right)=P\left(A\cap B\right)\end{aligned} \tag{3}$$

其中 M 为样本集；n 为样本集中属性的类别数目；d 为在样本集 M 中的 n 个类别属性中所占的比率。$I(M)$ 也被称作 M 的熵。

属性 A 对 M 的期望熵增益为 $SR(A)$，用增益率则可以表示为：

$$SR\left(A\right)=\frac{SR\left(A\right)}{Split\left(I\right)} \tag{4}$$

其中，$split\left(I\right)=\sum_{i=1}^{n}d\log_2 d$ 。

决策树的根用 $I(M)$最大值的属性，因此课堂测试成绩成为决策树的根节点，然后利用各个属性作为分支，生成决策树，图 3 展示的是决策树的一部分。

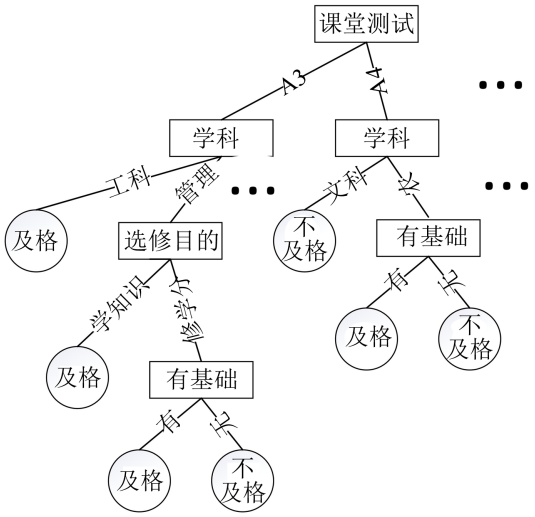

Figure 3. Diagram of the partial decision tree of C4.5 algorithm
图 3. C4.5 算法的部分决策树图

结果分析：

当课堂测试能够达到 A3 及以上，对于来自工科专业的学生期末成绩算法预测为及格；对于管理类专业的学生，如果学生的选课目的为学知识，那么算法预测为及格，对于已修学分为目的的管理类专业的学生，如果对于本课程有一定的基础或者人才培养方案中有相关的前置课程，那么算法预测为及格，否则算法预测为不及格。

当课堂成绩为 A4 的学生，学生如果来自文科专业，算法预测期末成绩为不及格；如果学生来自于农科专业，如果对于本课程有一定的基础或者人才培养方案中有相关的前置课程，那么算法预测为及格，否则算法预测为不及格。

不难发现，课堂测试对期末成绩的影响是非常大的，它不仅反应了对阶段性知识的学习效果，更是预测期末成绩是否通过的重要因素；当然为了更加细化和提高预测的精度，通过大量数据分析对于不同学科专业选修的学生，成绩表现还是有差异和规律的。

同时，选修目的的差异、有无相关基础知识的差异、培养计划中是否安排与本课程相关的基础课程并选修完毕等因素在很大程度上共同影响和决定了算法对于期末考试的预测效果。

通过对于期末成绩的预测，在很大程度上完成了“课程预警”最重要的一部分，从而为学生和授课教师从不同角度和层面提供了辅助支持。

6. 结论

通过对于“计算机网络与技术”通识课程的平台数据的关联分析，得到了较高支持度的规则 5 条，从这些规则中可以看出，网络学习平台对学生的自主学习起到辅助和指导的作用，但是也存在监督不足的特征，配合课堂测试这样的客观考察点进行联合分析，能够为教师教学安排和知识点讲授、任务点安排等方面提供具体的指导建议。

利用 C4.5 算法，综合考虑了课堂测试情况、选课学生的学科专业情况、专业人才培养方案情况、学生基础知识掌握情况等因素，构建了“学业预警”决策树，从而实现了本课程的期末考试成绩的预测，

对学生保持正常心理情绪、顺利毕业、教师教学安排等方面提供了支撑。

基金项目

本文受黑龙江省高等教育教学改革一般项目“‘一流本科’背景下通识课程改革研究与实践——以《计算机网络技术与用》为例”(编号：SJGY20210041)；东北林业大学教育教学研究项目“双一流”背景下后疫情时代研究生招生考试中的综合评价体系的研究与探索——以信息学院为例(DGYYJ2021-17)支持。

参考文献

[1] 张秀芹. 美国杜克大学通识教育的发展、特点及其借鉴意义[J]. 中国高等教育, 2020(5): 62-64.

[2] 吴凡, 陈诗敏, 赵泽宁. 大学生学习投入、学习时间及学习效果的比较研究——基于 F 省高校大学生线上线下学习经验调查[J]. 中国高教研究, 2022(10): 22-27+34.

[3] 许德心. 关联规则挖掘算法的并行化及应用研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2019.

[4] Bao, F.G., Mao, L.H., Zhu, Y.L., *et al*. (2022) An Improved Evaluation Methodology for Mining Association Rules. *Axioms*, **11**, 17. https://doi.org/10.3390/axioms11010017

[5] 王婷. 频繁项集挖掘算法的研究与应用[D]: [硕士学位论文]. 太原: 中北大学, 2019.

[6] 董云薪, 林耿, 张清伟, 陈颖婷. 基于 Apriori 算法填充数据及改进相似度的推荐算法[J]. 计算机科学, 2022, 49(S2): 307-311.

[7] 陆丽娜, 陈亚萍, 魏恒义, 杨麦顺. 挖掘关联规则中 Apriori 算法的研究[J]. 小型微型计算机系统, 2000(9): 940-943.

[8] 魏茜. 大类培养下校院两级学业预警指标体系的构建研究[J]. 黑龙江高教研究, 2020, 38(5): 55-59.

[9] Prabakaran, S., Ramar, R., Hussain, I., *et al*. (2022) Predicting Attack Pattern via Machine Learning by Exploiting Stateful Firewall as Virtual Network Function in an SDN Network. *Sensors*, **22**, 709. https://doi.org/10.3390/s22030709