文章编号：2095-3747（2022）-04-0039-05

基于关联规则的高职院校公共基础课程成绩分析与研究

庄群

（青岛远洋船员职业学院职教分院，山东 青岛 266427）

摘　要：实践教学中发现，不同公共基础课程成绩的优秀之间存在一定的关联规则。本文以某高职院校 11 个专业的学生公共课成绩为研究对象，提出了基于改进优秀率和数据整合的 D - Apriori 算法。新算法解决了由于课程不同评分标准引起的课程优秀成绩关联的偏离问题，能够更高效地探索公共基础课程之间获得优秀成绩之间的关联规则。相关研究结果能够为高职院校基础课程教学设计和教学方法的调整提供参考，为学生后续的培养和教育提供参考和建议。

关键词：高职院校；D - Apriori 算法；优秀率；数据挖掘

中图分类号：TP311　　文献标识码：A

职业教育中的公共基础课程能够为学生树立良好的思想品德、掌握专业知识、发展相关技能打下夯实的基础。[1] 实践教学中发现，不同公共基础课程成绩的优秀之间存在一定的关联规则。挖掘公共课程成绩优秀之间的关联规则，可以帮助老师在教学过程中更好地进行教学设计和调整教学方法。由于公共课程之间存在着不同的评分标准，从而会引起课程优秀成绩关联的偏离问题。因此，如何从不同评分标准的课程成绩中挖掘其真实的关联关系，为学生后续的培养和教育提供参考和建议具有重要的意义。

1 研究背景

目前，基于 Apriori 规则的数据挖掘算法被广泛应用于高校数据挖掘中。郭鹏等用改进的 Apriori 算法进行学生成绩的挖掘与分析，得到课程间关联规则 [2]；李绍中基于 Apriori 进行学生成绩分析找出学习态度、成绩相近的类别 [3]；张甜等对学生高考成绩与大学成绩进行数据挖掘，发掘学生入学前成绩对入学后表现关联规则 [4]。以上论文在数据分析和挖掘过程中未考虑到不同课程的评分特点。

本文以某航海院校近三年学习过高数、英语、毛概、体育、计算机的 11 个专业的学生成绩为数据研究对象，采用基于改进优秀率和数据整合的 D-Apriori 算法进行数据挖掘，不按照分数高低作为判定优秀的标准，而是按照学生成绩在某门课程中的排名判断学生成绩是否优秀。这就解决了由于基础课程成绩的不同评分特点对课程成绩优秀关联关系的偏离影响。同时，该算法对优秀率统一设置并进行数据离散化处理，更高效地探索公共基础课程之间获得优秀成绩之间的关联规则。结果表明，基础课程获得优秀成绩存在相关性，特别是数学课程优秀，在其它课程中获得优秀的可能性更高。数据挖掘的结果能够为有针对性地调整相关师资和教学方法，为学生后续的培养和教育提供参考和建议，提高基础课程的教学质量，提升学生的学习效果。[5-7]

2 基于 D - Apriori 算法的数据挖掘

2.1 传统 Apriori 算法的数据挖掘

Apriori 算法采用逐层搜索的迭代方法，其

收稿日期：2022—08—01

作者简介：庄群，女，硕士，讲师

基金项目：山东省职业教育教学改革研究项目：“信息技术环境下高职数学课程知识体系研究与实践”（项目编号：2017646）

中 k 项集用于探索 (k+1) 项集。其思路为：首先，通过扫描数据库，累计每个项的计数，并收集满足最小支持度的项，找出频繁 1 项集的集合。该集合记为 L1；然后，使用 L1 找出频繁 2 项集的集合 L2，使用 L2 找出 L3，如此下去，直到不能再找到频繁 k 项集；每找出一个 Lk 需要一次数据库的完整扫描。Apriori 算法基于最小支持度和置信度来探索关联关系，满足最小支持度和最小置信度的关联规则，即待挖掘的关联规则。

支持度:

$$\text{Support}(X => Y)=\frac{count(X\&Y)}{count\text{All}}$$　　公式（1）

置信度：

$$\text{Confidence}(X => Y)=\frac{count(X\&Y)}{count\text{A}}$$　　公式（2）

传统 Apriori 算法在进行学生成绩的数据挖掘时，未考虑不同课程的考核特点，存在一定缺陷。如表 1 所示，计算机课程成绩老师打分相对偏高，计算机课程的优秀率 41.74% 偏高（大学成绩通用以“>85 分”作为优秀判定标准），利用 Apriori 算法进行数据挖掘，会导致其他基础课程与计算机课程优秀成绩关联的偏离。

表 1　学生基础课成绩分布特征

汇总项	英语	毛概	高数	体育	计算机
平均值	70.34	72.05	68.53	74.20	79.90
最大值	95.5	96.00	100.00	97.5	100.00
最小值	31.5	30.00	15.00	46.00	21.00
优秀率（%）	7.88	5.70	13.76	6.17	41.74

为了避免不同基础课程成绩的评分特点影响对不同课程优秀成绩的探索，本文利用基于改进优秀率和数据整合的 Apriori 算法进行数据挖掘。

2.2 基于 D-Apriori 算法的优秀率选择

为降低课程优秀成绩关联的偏离，本文改进了优秀率和数据整合，不按照分数而是按照学生成绩在某门课程中的排名作为判定优秀的标准。如表 2 所示，对比各学生基础课成绩前 10% 和前 15% 分布，我们可以看出各基础课程前 15% 左右的成绩相对差别不大，且以排名 15% 作为优秀判定标准相比排名前 10% 作为优秀数据分析的每门课程优秀数量多 50%，更有利于探究各课程优秀率的关联规则。因此本文以课程成绩排名在前 15% 的认定为优秀，课程成绩排名不在前 15% 认为不优秀。

表 2　学生基础课成绩 10% 和 15% 分布比较

汇总项	英语	毛概	高数	体育	计算机
第 10%分数	84	82	88	82.5	92
第 15%分数	81	80.5	83	81	86
平均值	70.34	72.05	68.53	74.20	79.90

如下表 3 所示。相较于“课程成绩≥ 85”作为判定标准，“课程前 15% 的成绩”作为优秀的判断标准，各课程的优秀率不会出现过高（如计算机≥ 85 的占 41.74%）或过低（如毛概≥ 85 的仅占 5.60%）的问题；同时，能够避免由于老师打分特点造成的对课程成绩优秀的影响。此外，以“课程前 15% 的成绩”作为优秀的判断标准，每门基础课程的优秀率相同（都为 15%）。

表 3　学生成绩优秀率对比

优秀标准 / 科目	≥85		前 15%	
	分值	优秀率(%)	分值	优秀率(%)
英语	85.00	7.88	81.36	15.00%
毛概	85.00	5.70%	80.50	15.00%
高数	85.00	13.76%	83.00	15.00%
体育	85.00	6.17%	81.00	15.00%
计算机	85.00	41.74%	92.00	15.00%

2.3 D-Apriori 算法流程

利用 D-Apriori 算法进行数据挖掘，流程如图 1 所示：

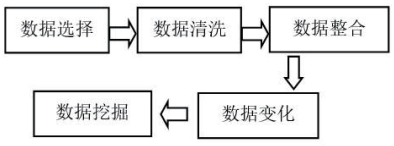


图 1　数据挖掘流程图

（1）数据选择：根据研究需要选择某高职院校 2016-2019 年学生基础课程成绩作为研究对象。

（2）数据清洗：去除研究数据中的无效数据。

（3）数据整合：数据中存在同一基础课程分布在不同学期的情况，将存在多个学期修读的课程按成绩学分加权取平均值。

（4）数据变化：对数据进行离散化，以学生成绩排名为依据，判断学生成绩是否优秀，优秀标记为 1；非优秀标记为 0。

（5）数据挖掘：编码，进行数据挖掘，获得相关输出结果。

基于 D-Apriori 算法具体数据流程过程的算

法流程图如图 2 所示：

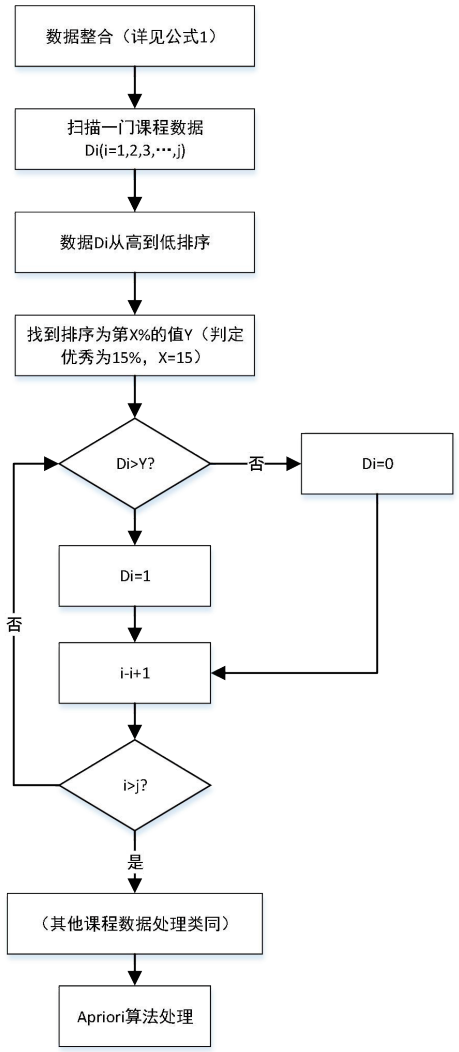


图 2　D-Apriori 算法流程图

2.4 基于改进优秀率和数据整合的 D–Apriori 算法的挖掘过程

（1）数据选择

选择 2016–2019 级基础课程中包含高数、英语、毛概、体育、计算机的 11 个专业的学生成绩为研究对象。学生成绩如表 4 所示。

表 4 学生成绩表

学号	姓名	计算机基础与应用	高等数学	体育（一）	体育（二）	大学英语阅读一	…
**01	***	65.00	70.00	98.00	74.00	64.00	…
**02	***	40.00	41.00	84.00	61.00	62.00	…
**04	***	85.00	60.00	98.00	76.00	74.00	…
**05	***	84.00	60.00	80.00	64.00	73.00	…
**06	***	64.00	60.00	80.00	52.00	61.00	…
…	…	…	…	…	…	…	…

（2）数据清洗

去除学生成绩中无效数据，如作弊、缺考或课程成绩缺失等无效数据记录。

（3）数据整合

记录中存在同一基础课程在不同学期有多个课程，例如体育课由体育（一）、体育（二）分两个学期修完。将存在多个学期修读的课程按成绩学分加权取平均值。

$$D=\sum_{k=1}^{n} a_k d_k / \sum_{k=1}^{n} a_k$$　（公式 3）

d_k（k=1,2,3……）为每学期课程成绩；

a_k 为对应学期课程成绩的学分。

（4）数据变化

为提高数据挖掘的效率，进行数据离散化处理。根据特定标准（每门课程成绩排名前 15%）判断学生成绩是否优秀。优秀标记为 1，非优秀标记为 0。具体流程如下：① 对某门基础课程成绩进行排序；② 根据成绩数据总量找出排名在 15% 的成绩 a;③ 将每个成绩与 a 进行比较，如果大于等于 a, 则赋值为 1，否则赋值为 0。④ 其他基础课程成绩变化按照流程 ①②③ 处理。数据离散化后如表 5 所示：

表 5　数据变化

序号	I1	I2	I3	I4	I5
T001	0	0	0	0	1
T002	1	0	0	0	0
T003	1	1	0	0	0
T004	0	0	1	1	0
T005	0	0	0	1	1
T006	0	1	0	0	1
T007	1	0	1	0	0
……	…	…	…	…	…

（5）数据挖掘

设定最小置信度为 15%，最小支持度为 2.25%，进行编码 , 运行并输出结果详见表 7。

3 数据挖掘结果与分析

分别对传统 Apriori 算法及改进优秀率和数据整合的 D–Apriori 算法的数据进行数据挖掘。设定最小置信度为 15%，最小支持度为 2.25%。基于传统 Apriori 算法进行数据挖掘结果如表 6 所示。

表 6　基于 Apriori 算法进行数据挖掘结果

序号	规则前项	规则后项	支持度	置信度
1	数学=优秀	英语=优秀	4.12%	29.91%
2	数学=优秀	计算机=优秀	8.11%	58.97%
3	英语=优秀	计算机=优秀	5.00%	63.43%
4	体育=优秀	计算机=优秀	3.06%	49.52%
5	毛概=优秀	计算机=优秀	3.35%	58.76%
6	英语=优秀	数学=优秀	4.12%	52.24%

表 7　基于 D-Apriori 数据挖掘结果

序号	规则前项	规则后项	支持度	置信度
1	数学=优秀	英语=优秀	4.59%	30.59%
2	数学=优秀	体育=优秀	6.35%	42.35%
3	数学=优秀	毛概=优秀	2.71%	18.04%
4	数学=优秀	计算机=优秀	4.41%	29.41%
5	英语=优秀	体育=优秀	4.00%	26.67%
6	英语=优秀	概论=优秀	3.00%	20.00%
7	英语=优秀	计算机=优秀	3.59%	23.92%
8	体育=优秀	概论=优秀	2.88%	19.22%
9	体育=优秀	计算机=优秀	3.76%	25.10%
10	毛概=优秀	计算机=优秀	2.71%	18.04%
11	英语=优秀	数学=优秀	4.59%	30.59%
12	体育=优秀	数学=优秀	6.35%	42.35%
13	毛概=优秀	数学=优秀	2.71%	18.04%
14	计算机=优秀	数学=优秀	4.41%	29.41%
15	体育=优秀	英语=优秀	4.00%	26.67%
16	毛概=优秀	英语=优秀	3.00%	20.00%
17	计算机=优秀	英语=优秀	3.59%	23.92%
18	毛概=优秀	体育=优秀	2.88%	19.22%
19	计算机=优秀	体育=优秀	3.76%	25.10%
20	计算机=优秀	毛概=优秀	2.71%	18.04%

通过表 7 基于改进优秀率和数据整合的 D-Apriori 算法数据挖掘结果可以发现：

① 基于改进优秀率和数据整合的 Apriori 算法由于提前设置了每门课程的优秀率都为 15%，前项和后项对换支持度和置信度不变（如表 7 中第 1 条关联规则与第 11 条的对比等）。基于改进优秀率和数据整合的 D-Apriori 算法，我们共获得了 20 条关联规则。

② 从整体看，各基础课程优秀相关的置信度 >15%, 高于每门课程的优秀率，验证了“高职学生的基础课程的学习成绩存在一定相关性，学生在一门基础课程成绩优秀，在其它课程的学习中取得优秀的可能性比其他同学更大。”这个发现。

③ 当设置支持度为 4% 时，我们发现数学成绩优秀的同学在英语、体育和计算机课程上获得优秀的可能性相对较高，英语成绩优秀的同学，体育课上获得优秀的可能性较高；反之亦然（基于 ①，对换规则前后项，结果相同）。

④ 基于传统 Apriori 算法进行数据挖掘，会导致基础课程优秀成绩关联的偏离。例如，未改进计算机课程的优秀率 41.74% 偏高（如表 3 所示），当以计算机为规则后项时，会导致置信度偏高（如表 6 所示）

综上，D-Apriori 数据挖掘能够降低由于基础课程成绩的不同评分特点对课程成绩优秀关联关系的影响。这种方法能够更加真实、高效地探索公共基础课程之间获得优秀成绩之间的关联规则。

4 结论

基于改进优秀率和数据整合的 D-Apriori 算法能够简单高效地探索公共基础课程之间获得优秀成绩之间的关联规则，降低由于基础课程成绩的不同评分特点对课程成绩优秀关联关系的影响。数据挖掘的结果发现：数学基础课程优秀，在其他课程中获得优秀的可能性更高，良好的数学基础有助于促进学生其他课程的学习。高职院校在今后的基础教学过程中，可以适当增加数学课程的学时，有针对性地调整教学方法，注重培养学生对数学课程学习的兴趣，激励学生自主探究和学习基础课程问题的能力，从而促进其他基础课程知识与思维能力的提升，实现提高学生的整体成绩的目标。

本文存在一定局限性，由于各地区高职院校存在一定行业特色和不同高校中学生文理科比例不同，每个高职院校的基础课程成绩存在不同，相关数据研究结果还需要补充其他样本。

参考文献：

[1] 周庆 , 牟超 , 杨丹 . 教育数据挖掘研究进展综述 [J]. 软件学报 ,2015,26(11):3026-3042.
[2] 郭鹏 , 蔡骋 . 基于聚类和关联算法的学生成绩挖掘与分析 [J]. 计算机工程与应用 ,2019,55(17):169-179.
[3] 李绍中 . 数据挖掘改进算法在学生成绩分析中的应用 [J]. 科技通报 ,2012,(08)：216-217,220.

[4] 张甜 , 尹长川 , 潘林 , 安杰 . 基于改进的聚类和关联规则挖掘的学生成绩分析 [J]. 北京邮电大学学报 (社会科学版),2018,(02):91–96.
[5] 薛南 , 吕柏权 , 倪陈龙 . 基于自编码器和填充函数的深度学习优化算法 [J]. 电子测量技术 ,2019,(23):79–84.
[6] 袁路妍 , 李锋 . 改进的关联规则 Apriori 算法在课程成绩分析中的应用 [J]. 中国教育信息化 , 2017,(17):62–64.
[7] 徐承俊 , 朱国宾 . 数据挖掘在全国计算机等级考试 (NCRE) 成绩分析中的研究及应用 [J]. 计算机应用与软件 , 2020,(08):64–67,73.
[8] 张逸谦 , 王志强 , 梁正平 . 基于托普利兹逆协方差的学习状态挖掘算法 [J]. 现代教育技术 ,2021,(08):119–126.
[9] 贾国栋 . 多相关周期性时间序列上的异常模式关联规则挖掘 [D]. 东北大学 ,2010:17–18.

Performance Mining and Analysis of Higher Vocational Colleges based on Apriori Optimization Algorithm

ZHUANG Qun

（School of Vocational Education，Qingdao Ocean Shipping Mariners College，Qingdao 266427，China）

Abstract: In practical teaching, it is found that students' academic performance in basic courses is correlated to some extent. This paper takes the students' public course scores of 11 majors in a higher vocational college as the research object, and proposes a D–Apriori algorithm based on improved excellence rate and data integration. The new algorithm solves the problem of deviation in the association of excellent grades caused by different scoring standards of courses, and can more efficiently explore the association rules between excellent grades obtained by public basic courses. The relevant research results of this paper can provide reference for the teaching design of basic courses and the adjustment of teaching methods in higher vocational colleges, and provide reference and suggestions for the follow–up training and education of students.

Keywords: vocational college; apriori algorithm; excellence rate;data mining.