

关联规则算法在教学评价数据分析中的应用

——以Y大学为例

简相栋, 王文东*, 甄艳秋

延安大学数学与计算机科学学院, 陕西 延安

收稿日期: 2021年10月25日; 录用日期: 2021年11月23日; 发布日期: 2021年11月30日

摘要

数据库技术越来越成熟, 应用非常广泛, 存储的数据量越来越大。大多数数据往往只是简单处理和应用, 并没有进行数据深度挖掘, 造成数据资源的浪费。关联规则算法可以有效地挖掘出数据中不同事务之间的微小关系, 教学评价中, 教师职称、教师年龄、教师教学时长与教学态度、教学纪律、教学方法、教学效果、师德教风等方面密切相关。通过apriori算法和Fp-growth算法两个典型关联规则算法的分析研究, 详细叙述两种算法的题解过程, 利用两种算法来研究Y大学的教学评价数据, 发现高职称年长教师的授课效果较好, 但是学生的认可度较低, 年轻教师的授课效果不太好, 学生的认可度反而较高, 说明老教师在沟通交流方面不满足青年学生的期望, 青年教师的教学能力和教学方法有待提升。

关键词

Apriori算法, 关联规则, 教学评价, 高校教育

Application of Association Rule Algorithm in Data Analysis of Teaching Evaluation

—Taking University Y as an Example

Xiangdong Jian, Wendong Wang*, Yanqiu Zeng

School of Mathematics and Computer Science, Yan'an University, Yan'an Shaanxi

Received: Oct. 25th, 2021; accepted: Nov. 23rd, 2021; published: Nov. 30th, 2021

Abstract

Database technology is becoming more and more mature, widely used, and the amount of data stored is increasing. Most data are often simply processed and applied without deep data mining,

*通讯作者。

resulting in a waste of data resources. Association rule algorithm can effectively mine the small relationship between different transactions in the data. In teaching evaluation, teachers' professional title, teachers' age and teachers' teaching time are closely related to teaching attitude, teaching discipline, teaching methods, teaching effect, teachers' ethics and teaching style. Through the analysis and research of two typical association rule algorithms, apriori algorithm and FP-growth algorithm, the problem solving process of the two algorithms is described in detail. The two algorithms are used to study the teaching evaluation data of Y University. It is found that the teaching effect of senior teachers with high professional titles is better, but the recognition of students is low, the teaching effect of young teachers is not very good, and the recognition of students is higher, it shows that the old teachers do not meet the expectations of young students in communication, and the teaching ability and teaching methods of young teachers need to be improved.

Keywords

Apriori Algorithm, Association Rules, Teaching Evaluation, College Education

Copyright © 2021 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. Apriori 算法

关联规则是 1993 年, 由 R. Agrawal 等人创建的。主要研究的是事物内部属性之间的关联性。美国的沃尔玛超市曾经利用过关联规则, 研究“啤酒与尿布”之间的关系, 发现消费者在购买啤酒的时候, 大概率会买尿布, 于是沃尔玛超市就把啤酒和尿布放在了一起, 从而提高了营业额。

1.1. 基本概念

项与项集: 设 $itemset = \{item_1, item_2, \dots, item_m\}$ 是所有项的集合, 其中, $item_k (k = 1, 2, \dots, m)$ 成为项。项的集合称为项集(itemset), 包含 k 个项的项集称为 k 项集(k -itemset) [1]。

事务与事务集: 一个事务 T 是一个项集, 它是 itemset 的一个子集, 每个事务均与一个唯一标识符 Tid 相联系。不同的事务一起组成了事务集 D , 它构成了关联规则发现的事务数据库。

定义. 设 $I = \{I_1, I_2, I_3, \dots, I_m\}$ 是项的集合, 其中 m 表示项的数目。事务数据库 $D = \{t_1, t_2, \dots, t_m\}$, $t_i \in I$, $T \neq \emptyset$ 。

支持度(support): 是两个事物同时出现的概率, 如果两者同时出现的概率比较大, 则说明两者是有相关性的, 若两者同时出现的概率小, 则关系不大。按数学定义, 集合 Z 在事物集 D 中支持度表示在事物集 D 中任取一个集合, 集合中包含 z 的概率。即:

$$\text{支持度}(\text{support})(Z) = \frac{|\{T | T \in D \text{ and } Z \subset T\}|}{|D|},$$

关联规则 $X \Rightarrow Y$ 支持度表示为:

$$\text{Support}(Z \Rightarrow Y) = \text{Support}(Z \cup Y) = \frac{|\{T | T \in D \text{ and } (Z \cup Y) \subset T\}|}{|D|}, \text{ 其中“}|\{|\text{”表示集合个数。}$$

置信度(confidence): 数学定义里面, 置信度是两者同时发生或者出现的概率, 相当于条件概率。

$$\text{Confidence}(Z \Rightarrow Y) = \frac{\text{support}(Z \cup Y)}{\text{support}(Z)} = \frac{|\{T | T \in D \text{ and } (Z \cup Y) \subset T\}|}{|\{T | T \in D \text{ and } X \subset T\}|}$$

提升度[2]: 提升度是为了判断两个元素的关联程度有多大, 即支持集合 Z 对支持集合 Y 的影响有多大, 提升度大于 1 且越高, 正相关性越高。提升度小于 1 且越低则相反。提升度在等于 1 时表示两个是独立的个体。数学表达式如下:

$$\text{Lift}(Z \Rightarrow Y) = \frac{P(Y|Z)}{P(Z)}$$

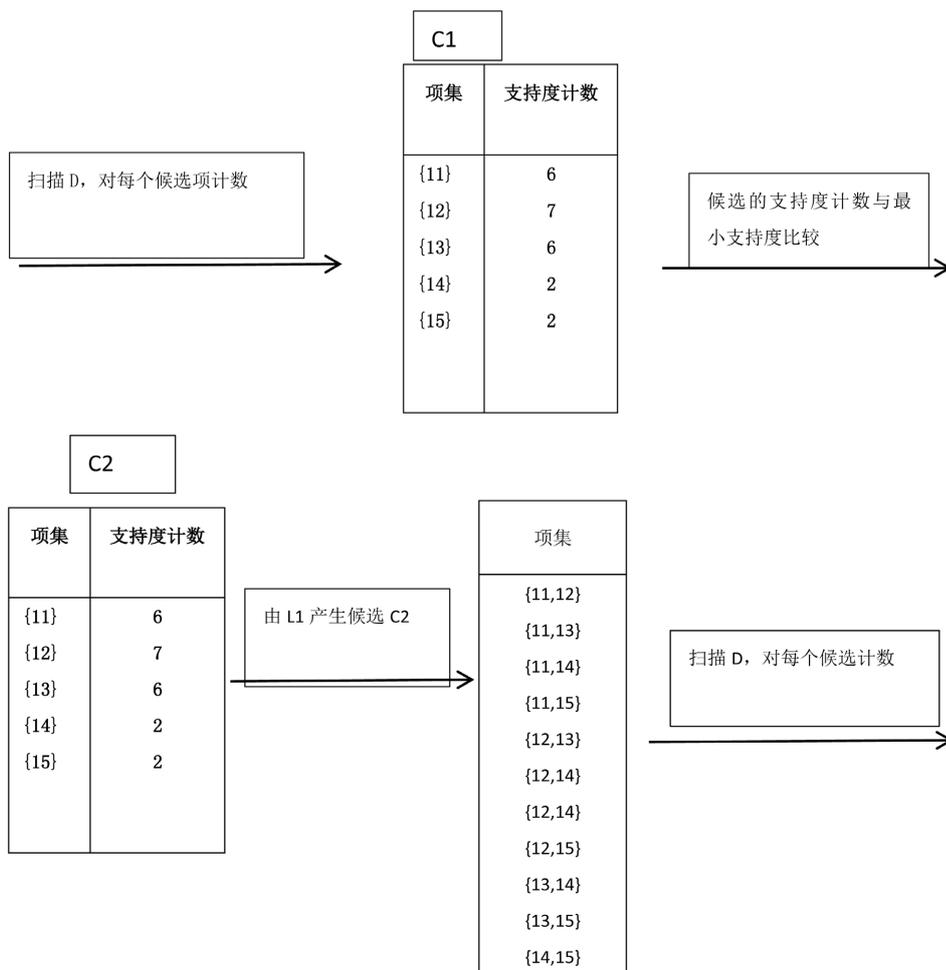
我们经常会使用上面这三个指标, 来进行关联规则的衡量, 筛选出符合条件的关联规则。满足最小的支持度和可信度的规则, 被称为强关联规则, 提升度大于 1 时, 表示有效的强关联规则。提升度小于 1 时, 表示无效的强关联规则。提升度等于 1 时表示两个是独立的个体。

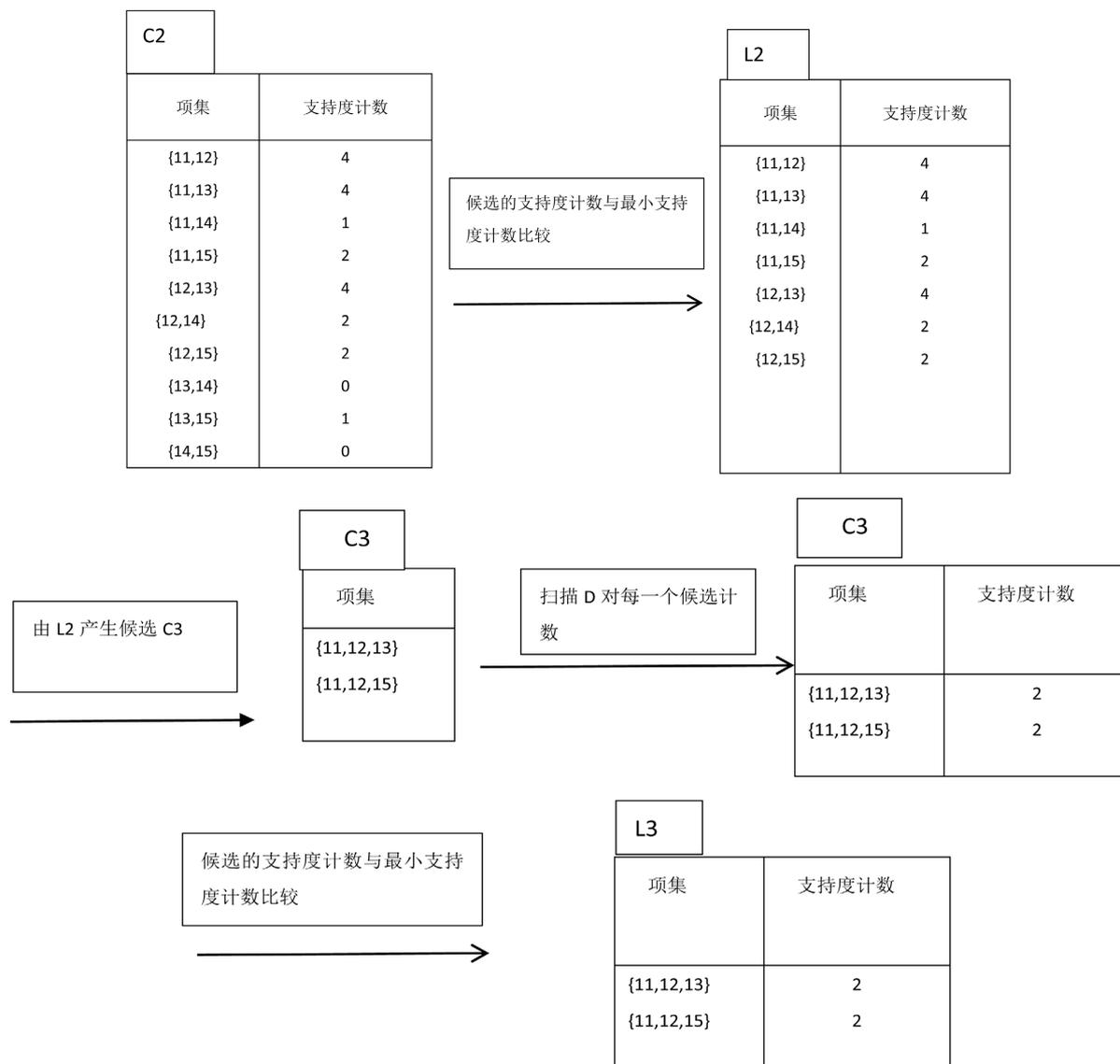
项集的出现频度(support count): 包含项集的事务数, 简称为项集的频度、支持度计数或计数。

频繁项集(frequent itemset): 如果项集 I 的相对支持度满足事先定义好的最小支持度阈(即 I 的出现频度大于相应的最小出现频度(支持度计数)阈值), 则 I 是频繁项集。

强关联规则: 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。

1.2. 图解过程





1.3. 核心代码

```

L1 = find_frequent_1-itemsets(D);
for (k=2; Lk-1 ≠ ∅ ; k++) {
    Ck = apriori_gen(Lk-1, min_sup);
    for each transaction t ∈ D //scan D for counts
        Ct = subset(Ck,t); //get the subsets of t that are candidates
        for each candidate c ∈ Ct
            c.count++;
        }
    Lk = {c ∈ Ck | c.count ≥ min_sup}
}
return L = ∪ k Lk;
    
```

2. Apriori 算法在教学评价数据分析中的应用

2.1. 介绍数据来源

为了提高教学质量, 全面了解课堂教学情况, 及时反馈教学信息, 绝大多数的高等院校, 每年都会安排很多次的调查问卷, 安排学生来评价代课老师的教学质量。本次数据来源于延安大学本科生, 各专业近三年的教学评价数据。将很多指标总结性归纳为, 教学态度, 教学内容, 教学方法, 教学纪律, 教学表达, 总体效果。具体如表 1 教学评价项目表:

Table 1. Teaching indicators

表 1. 教学指标

评价项目	评价内容
教学态度	教书育人态度, 课堂授课态度, 吸取别人意见的态度, 治学态度等。
教学内容	教学内容的系统性、综合性, 信息量, 深浅程度的把握, 理论与实际的联系, 学科前沿知识的介绍, 基础理论知识、重点难点知识的传授与技能的培养。
教学方法	适合学科特点和学生特点, 对学生学习方法的指导, 因材施教, 师生互动, 启发学生积极思维, 学生能力的培养, 现代教学手段的运用, 教学改革等。
教学纪律	学生上课的精神及注意力, 教师对课堂教学秩序的管理。
教学表达	语言的规范化, 口头语言表达能力, 体态语言的运用, 板书设计等。
总体效果	讲课的吸引力, 课堂气氛, 学生对基础知识、理论、技能的掌握, 分析和解决问题能力的培养等。

2.2. 数据处理

教学系统导出数据, 一般情况下是不完整的, 需要数据清洗和整理, 并采用加权平均的方法计算每个测评项目的分值。每个项目的取值如下:

- 1) 年龄: 老, 中, 青
- 2) 职称: 教授, 副教授, 讲师, 助教
- 3) 教学态度: 优, 良, 中, 差
- 4) 教学内容: 优, 良, 中, 差
- 5) 教学方法: 优, 良, 中, 差
- 6) 教学纪律: 优, 良, 中, 差
- 7) 教学表达: 优, 良, 中, 差
- 8) 总体效果: 优, 良, 中, 差

离散化处理后, 教学课堂部分数据见表 2。

Table 2. The teaching classroom part data

表 2. 教学课堂部分数据

工号	年龄	职称	教学态度	教学内容	教学方法	教学纪律	教学表达	教学效果
16020***	青年	硕士研究生	优	优	中	优	中	优
19060***	中年	博士研究生	中	优	优	优	优	优

Continued

13010***	青年	硕士研究生	优	中	优	优	中	优
16010***	青年	博士研究生	优	优	优	优	中	优
13010***	中年	大学本科	中	优	优	优	优	中
15010***	青年	硕士研究生	优	中	优	优	优	优
13010***	中年	大学本科	优	优	中	优	优	优
13010***	青年	硕士研究生	中	优	优	中	优	优
13010***	青年	硕士研究生	优	优	中	优	优	优
13020***	中年	专科及以下	优	中	优	优	优	中
13020***	中年	专科及以下	中	优	优	优	优	中
13010***	青年	硕士研究生	优	优	优	优	优	优
13010***	青年	硕士研究生	优	优	优	优	优	优
13010***	青年	硕士研究生	优	优	优	优	优	优
13010***	中年	大学本科	优	中	优	优	优	中
14020***	青年	硕士研究生	优	优	优	优	优	优
19010***	青年	博士研究生	中	优	中	优	优	优
13020***	青年	硕士研究生	优	中	优	优	优	优
13010***	青年	博士研究生	优	优	中	优	优	中
.....

2.3. 数据结果分析[3]

单个指标对教学结果的影响

年龄与整体效果的关联性如表 3 表示:

Table 3. The association between age and overall effect

表 3. 年龄与整体效果的关联性

前项	后项	支持度%	置信度%
总体效果 = 优	年龄 = 青年	38.9	43.82
总体效果 = 优	年龄 = 中年	47.8	50.5
总体效果 = 优	年龄 = 老年	35.6	68.5
.....

教学方法与总体效果的关联性如表 4 表示[4]:

Table 4. The relevance of teaching methods to overall effectiveness**表 4.** 教学方法与总体效果的关联性

前项	后项	支持度%	置信度%
总体效果 = 优	教学方法 = 优	37.5	46.98
总体效果 = 良	教学方法 = 中	30.5	45.24
总体效果 = 中	教学方法 = 中	20.4	28.3
.....

职称与总体效果的关联性如表 5 表示:

Table 5. The relevance of the title to the overall effect**表 5.** 职称与总体效果的关联性

前项	后项	支持度%	置信度%
总体效果 = 优	职称 = 教授	29.52	46.49
总体效果 = 良	职称 = 讲师	38.9	26.5
总体效果 = 优	职称 = 副教授	30.1	46.24
总体效果 = 中	职称 = 讲师	37.8	34.7
.....

实验结果表明[5]:

1) 只考虑职称的因素, 教授的讲课效果比较好, 但认可度不高, 可能和教授比较严厉有关系。讲师的讲课效果有待提高, 但认可度都比较高, 这可能和年轻教师比较有活力有一定的关系。

2) 只考虑教学方法的因素, 总体效果和教学方法的优良中是成正比例关系的, 教学方法比较优的, 对应的教学效果就好。

3) 只考虑年龄的因素, 年龄较大教师的授课效果好, 但认可度, 没有年轻教师好, 中年教师和青年教师的认可度相对高一些, 但授课效果欠缺[6]。

对于以上问题, 针对于 Y 大学而言, 老年教师的授课效果较好, 认可度低。年轻教师的授课效果不太好, 而认可度较高, 表明老教师和学生的交流较少, 或者说明老教师的上课的生动性不如年轻教师, 而年轻教师的教学效果较差, 需要多进行培训。

3. 总结

关联规则算法, 还有很多需要改进的地方, 比如产生很多的候选项集, 占用内存过大, 运算速度慢[7]。我们可以尝试从改进数学公式或者扫描方式来考虑优化[8], 也可以从存储方式来考虑, 或者我们是不是能够针对不能的数据形式或者研究方向来改进我们的算法[9]? Apriori 算法和 Fp-growth 算法是关联规则中的经典算法, 本文用于研究教学评价数据中的教学指标, 实验数据表明, 教学效果与年龄、职称和教学方法有很强的关联性, 具体表现在职称高的, 年纪大的, 教学效果稍微好一点, 但教师的认可度

低, 不受学生们的喜欢。年轻教师的教学效果差一些, 教学经验上没有老教师多, 但年轻教师的认可度高, 受学生们的喜欢。

基金项目

延安市科技局专项科研项目(2018KG-02); 陕西省教育厅教学研究与改革项目(19BZ031); 延安大学研究生教育创新计划项目(YCX2020101); 教育部高等学校计算机类专业教学指导委员会系统能力培养试点院校项目(2019-24)。

参考文献

- [1] 李竹林, 刘芬. 数据挖掘算法研究与实现[M]. 北京: 中国水利水电出版社, 2017.
- [2] 马瑞敏, 吴霞. 基于 FP-Growth 算法的关联规则挖掘研究与应用[J]. 太原师范学院学报, 2021, 20(1): 19-22.
- [3] 文继权, 屈武江. 关联规则挖掘在教学评价中的应用研究[J]. 计算机与现代化, 2010(8): 25-28.
- [4] 王峻. 关联规则在课堂教学测评中的应用[J]. 白城师范学院学报, 2018(10): 52-57.
- [5] 邓慧. 基于 Apriori 算法的教学评价系统[J]. 电脑知识与技术, 2013(10X): 6695-6697.
- [6] 陈庆. 关于大学本科教学管理的几点思考[J]. 科技风, 2019(32): 82.
- [7] 程一芳. 数据挖掘中的数据分类算法综述[J]. 技术分析, 2021(2): 136-140.
- [8] 韩存鸽. FP-growth 算法的研究与改进[J]. 龙岩学院学报, 2020, 38(5): 8-13.
- [9] 郝林倩. 基于关联规则的数据挖掘算法分析[J]. 太原学院学报, 2020, 38(3): 42-45.