

文章编号: 1004-4353(2015)02-0160-04

# 改进的 Apriori 算法在成绩分析中的应用研究

侯继文, 徐善针\*

( 延边大学工学院 计算机科学与技术学科 智能信息处理研究室, 吉林 延吉 133002 )

**摘要:** 针对经典 Apriori 算法会产生大量冗余规则的缺点,在两方面对算法进行了改进:一方面是对产生频繁项集方式的改进,使算法只产生包含目标项的频繁项集;另一方面是对产生规则方式的改进,使算法只产生关联后件中包含目标项的关联规则. Apriori 算法改进前后的对比表明:改进后的 Apriori 算法可以避免非目标规则的产生,使算法更符合成绩分析的需要,提高算法的执行效率. 将改进的 Apriori 算法应用于成绩分析中表明,改进后的算法能够挖掘出各门前导课程成绩对后续课程成绩的影响,因此可为教师制定有针对性的教学计划提供参考.

**关键词:** 关联规则; 改进的 Apriori 算法; 成绩分析

**中图分类号:** TP391.1      **文献标识码:** A

## Application and research of an improved Apriori algorithm in score analysis

HOU Jiwen, XU Shanzhen\*

( Intelligent Information Processing Lab., Dept. of Computer Science & Technology,  
College of Engineering, Yanbian University, Yanji 133002, China )

**Abstract:** The classical Apriori algorithm which produces a large number of redundant rules has been improved in two aspects. On the one hand, the way to generate frequent item sets has been improved to only produce frequent item sets that contain targeted item. On the other hand, the way to generate rules has been improved to only produce the rules of associated consequent which contain targeted item. The comparison between Apriori algorithm and improved Apriori algorithm indicates that the improved Apriori algorithm can avoid the generation of nontarget rules, meet the demand of score analysis better, and improve the execution efficiency. When being used in score analysis, the improved algorithm can dig out the influence to the database score from preceding courses score. As a result, it can provide a reference for teachers to develop targeted teaching plan.

**Key words:** association rules; improved Apriori algorithm; score analysis

随着我国教育信息化的推广,各大高校都建立了教务管理系统,且在教务管理系统中存储着大量的学生成绩数据. 如何有效地分析这些数据,得出各门课程成绩之间的关联关系和影响成绩的主要因素等有价值的信息,用以指导教学是教育工作者们关心的问题<sup>[1]</sup>.

关联规则 Apriori 算法是一种用于挖掘数据项间关联关系的经典算法<sup>[2]</sup>,将关联规则 Apriori 算法应用到高校学生成绩分析中,可以挖掘出隐藏在成绩数据中潜在的有价值的知识<sup>[3]</sup>. 使用经典 Apriori 算法挖掘分析成绩数据,在产生有价值的目标规则的同时,也会产生大量冗余规则<sup>[4]</sup>,

给成绩分析带来不便. 本文针对经典 Apriori 算法所存在的不足,提出了改进的 Apriori 算法,并通过成绩分析实例验证了该算法的有效性.

1 经典 Apriori 算法

1.1 Apriori 算法原理

Apriori 算法是用来挖掘数据库中各数据项之间的频繁模式的一种算法<sup>[5]</sup>,此算法使用频繁项集性质的先验知识,以逐层搜索的迭代方式来获得频繁项集<sup>[6]</sup>. 算法的具体描述如下:

算法的第一个步骤是从原始数据集中找出所有频繁项目集(frequent itemsets). 某一项集出现的频率表示为支持度(support),以一个包含事件 A 与事件 B 的 2 项集为例,当{A,B}的出现频次不小于事先设定好的最小支持度阈值时,{A,B}称为频繁项集. 首先找到长度为 1 的频繁 1 项集,之后算法会从频繁 1 项集中产生频繁 2 项集,以此类推,直到无法产生更长的频繁项集为止<sup>[6]</sup>.

算法的第二步是利用第一步中产生的频繁项集来生成关联规则<sup>[7]</sup>,若一条关联规则的置信度不小于事先设定的最小置信度阈值,则称此规则为强关联规则. 例如:经由高频  $k$ -项集{A,B}所产生的规则  $A \rightarrow B$ ,若置信度大于等于最小置信度阈值,则称  $A \rightarrow B$  为强关联规则<sup>[8]</sup>.

1.2 Apriori 算法在成绩分析中的不足之处

用 Apriori 关联规则算法挖掘成绩数据可以方便、快捷地产生出所有满足最小支持度、最小置信度的关联规则<sup>[9]</sup>,但其中并不是所有规则都是教师们需要的结果. 以挖掘前导课程对后续课程的影响为例,教师想要了解前一学期的课程对自己所教课程的影响,要得到的结果应该是“规则后件”包含其所教课程的规则. 例如,教数据库课程的教师希望得到的是类似“如果计算机导论成绩优秀则数据库成绩也优秀”这种形式的规则,但此时,经典 Apriori 算法所产生的规则不止包括这些目标规则,许多类似“如果 C 语言成绩优秀则高等数学成绩也优秀”这种描述前导课程之间关联的规则也会被挖掘出来. 一些并不需要的规则掺杂在目标规则中,这不仅浪费筛选目标规则的时间,同时也会降低 Apriori 算法的计算速度.

2 Apriori 算法的改进

2.1 改进 Apriori 算法

假设“目标规则后件”为  $T$ ,改进的思想是只产生包含  $T$  的频繁项集. 首先改进产生频繁项集的步骤,其目标是使产生的规则中包含  $T$ . 由产生规则的方式可知,能够产生规则的最小频繁项集是频繁 2 项集,因此需要添加一个控制条件,使产生的频繁 2 项集都包含  $T$ . 频繁  $k$  项集是由两个前  $k-2$  项相同的  $k-1$  项集连接产生的,只要所有频繁 2 项集都包含  $T$ ,那么后续的频繁项集也都会包含  $T$ ,这就保证了产生的关联规则中一定包含  $T$ .

其次是改进生成规则的步骤,以上改进虽然保证了规则中只包含  $T$ ,但  $T$  既可能出现在规则的后件中,也可能出现在规则的前件中,即可能产生“如果数据库成绩优秀则计算机导论成绩也优秀”这种形式的规则,这显然也不是教师们想要得到的规则. 因此,需要在产生规则时再添加一个控制条件,使关联规则后件中包含  $T$ . 一条关联规则的前件是由一个频繁项集的非空子集构成,后件则是频繁项集去掉该子集后产生的集合,本文将该集合表示为  $F$ . 改进的生成规则步骤中需要加一个条件以判断  $F$  是否包含  $T$ ,如果包含,则产生规则,否则不产生. 由此就可以去掉不必要的规则,提高 Apriori 算法的效率.

改进部分的代码如下:

1) 产生频繁 2 项集处的代码:

```
for (ItemSet is2 : F2) {  
    if (! (is2.contains(A))) {  
        deleteList.add(is2); // 不包含 T 的记录保存到 delList 中  
    }  
}
```

$F2.removeAll(deleteList)$ ; // 从频繁 2 项集中去掉所有不包含 A 的频繁 2 项集

2) 产生规则的代码:

```
remainSet =  $L_k$  - subSet; // 用 remainSet 表示关联规则后件  
if(remainSet.contains(T)){ // 判断 remainSet 即关联后件中是否包含 T
```

```
output the rule subSet=> remainSet
}
```

2.2 改进算法与原算法的对比分析

为保证产生较多规则,设置最小支持度为 0.01,通过不断提高最小置信度,来对比改进算法和经典算法在相同数据集下所产生的规则数,结果如图 1 所示.由图 1 可以看出,在产生规则数较多的情况下,改进算法明显优于经典算法.改进算法所产生的规则是以  $T$  为后件的关联规则,即目标规则,而经典算法是无差别地产生所有规则.由表 1 可知,随着置信度的提高,经典算法产生的冗余规则虽然逐渐减少,但仍较多,而改进算法只产生在成绩分析中所需要的关联规则,因此可节省大量时间.

表 2 为经典算法和改进算法的运行时间,图 2 为两种算法执行时间的对比图.从表 2 和图 2 中可以看出,改进算法的执行效率明显高于原始算法.

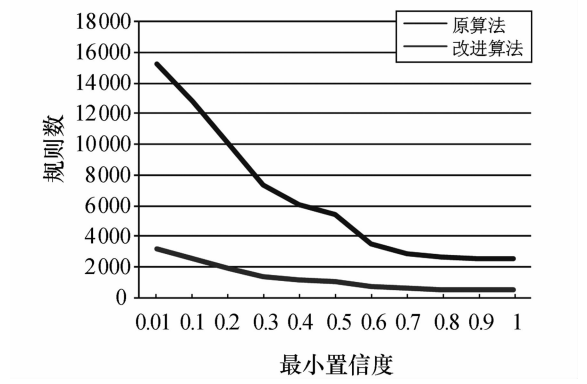


图 1 两算法产生的规则数对比

表 1 两算法产生的规则数

最小置信度	原算法规则数	改进算法规则数
0.1	12 795	2 629
0.2	10 009	1 971
0.3	7 390	1 422
0.4	6 039	1 183
0.5	5 469	1 070
0.6	3 538	730
0.7	2 871	628
0.8	2 648	580
0.9	2 544	552
1	2 540	550

表 2 两算法的运行时间

最小置信度	原算法运行时间/ms	改进算法运行时间/ms
0.1	1 206	1 064
0.2	1 094	923
0.3	974	883
0.4	970	798
0.5	963	751
0.6	892	689
0.7	879	673
0.8	829	661
0.9	828	658
1	827	658

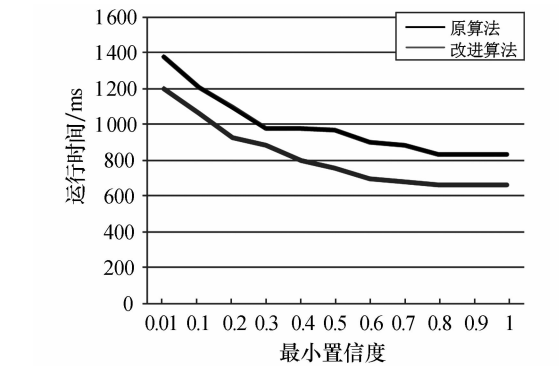


图 2 两算法运行时间的对比

3 基于改进 Apriori 算法的成绩分析

本文中的数据来源于延边大学教务管理系统中 2006—2010 级计算机专业期末考试成绩,去掉成绩缺失和缺考的记录后,共有 452 条成绩记录.为了便于数据挖掘,需对给定成绩进行预处理,成绩单包括系、学号、学院、计算机导论、C 语言、大学外语、大学物理、高等数学、离散数学等 12 个属性.首先将系、班级、学号这几个属性去掉,其余均用于挖掘分析的成绩属性,然后将成绩  $\geq 80$  分的离散化为优秀,  $70 \sim < 80$  为良好,  $60 \sim < 70$  为中等,低于 60 为较差.

挖掘的目标是分析出各门前导课程对后续课程数据库成绩的影响.设置目标后件  $T$  为“数据库”,目标规则为关联后件包含“数据库”的关联规则.设定最小支持度为 0.2,最小置信度为 0.7,然后执行改进的 Apriori 算法,得到的关联规则如下:

规则 1 如果[计算机导论=优秀,离散数学=优秀]则[数据库=优秀]置信度为 0.86.

规则 2 如果[计算机导论=优秀,C 语言=优秀]则[数据库=优秀]置信度为 0.84.

规则 3 如果[计算机导论=优秀,大学物理=优秀]则[数据库=优秀]置信度为 0.80.

规则 4 如果[计算机导论=优秀]则[数据库=优秀]置信度为 0.77.

规则 5 如果[离散数学=优秀]则[数据库=优秀]置信度为 0.75.

规则 1 表示如果计算机导论成绩优秀且离散数学成绩优秀,则数据库成绩也为优秀的概率为 86%,说明这 3 门课程间存在一定的关联,即计算机导论和离散数学成绩好的学生在学习数据库课程时会有一定优势.这是因为计算机导论作为计算机专业的基础学科,可为学习数据库打好基础,而离散数学中的逻辑运算与数据库中的逻辑运算相关联,因此学好离散数学对学习数据库也有一定帮助.同理可以由其他规则得到其他前导课程对数据库成绩的影响,教师可以根据这些信息,提前了解学生学习数据库课程的情况,从而有针对性地制定教学计划.

## 4 结论

本文根据成绩分析的特点,在产生频繁项集和生成关联规则两方面对 Apriori 算法进行了改进.通过添加控制条件的方式,使算法根据挖掘目标产生频繁项集,提高了算法的执行效率;通过控制条件过滤掉非目标规则方法,减少了算法的执行时间,也使得算法更符合成绩分析的需求.将改进

后的算法应用于高校学生成绩分析中,得出了理想的挖掘结果,可为教师优化教学方案提供参考.

本文在使用改进的 Apriori 算法进行成绩分析的过程中,仅分析了前导课程对后续课程的影响,分析角度较为单一.利用关联规则分析高考成绩、生源地对本科成绩的影响以及各题型对总分的影响等多角度的成绩分析是下一步的研究内容.

## 参考文献:

- [1] 广继红.数据挖掘在教务系统成绩分析中的应用研究[D].长春:吉林大学,2012:11-21.
- [2] 申义彩,杨枫,王晓燕.基于关联规则的计算机等级考试答卷分析[J].信息系统工程,2011,10:127-129.
- [3] 孙月昊.基于关联规则的成绩分析及课程设置研究[D].石家庄:河北科技大学,2013:18-19.
- [4] Buldu A, Üçgün K. Data mining application on students' data[J]. Procedia Social and Behavioral Sciences, 2010,2:5251-5259.
- [5] Nateka S, Zwillling M. Student data mining solution-knowledge management system related to higher education institutions[J]. Expert Systems with Applications, 2014,41:6400-6407.
- [6] Aher S B, Lobo L M R J. Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data[J]. Knowledge-Based Systems, 2013,51:1-4.
- [7] 王欣,徐腾飞,唐连章. SQL Server 2005 数据挖掘事例分析[M].北京:中国水利水电出版社,2008:128-131.
- [8] Lazcorreta E, Botella F, Caballero A F. Towards personalized recommendation by two-step modified Apriori data mining algorithm[J]. Expert Systems with Applications, 2008,15:1422-1429.
- [9] 耿悦杰.关联规则算法在教育信息数据挖掘中的应用[J].计算机与现代化,2012,5:83-85.