

文章编号: 1004-4353(2019)01-0040-05

# 基于兴趣度关联规则的在线学习行为分析方法

胡延雪, 怀丽波\*, 崔荣一

( 延边大学 工学院, 吉林 延吉 133002 )

**摘要:** 针对如何使用数据挖掘技术分析指导用户改善学习行为的问题,提出了一种基于兴趣度关联规则的学习行为分析方法.首先,采用  $K$ -means 聚类方法快速归纳出用户的学习状态;其次,通过含兴趣度的关联规则算法获得学习行为与学习效果之间的强规则;最后,以 edX 平台提供的用户学习数据为例对算法进行了验证.结果表明:含兴趣度指标的算法所获得的强规则数目比传统关联规则算法缩减了 40.9%,同时该方法能够得出学习行为因素与学习效果之间的具体关系,有利于指导用户改善学习行为.

**关键词:** 在线课堂; 学习行为; 聚类; 关联规则; 兴趣度

**中图分类号:** TP399

**文献标识码:** A

## Research on online learning behavior analysis method based on the association rule of degree of interest

HU Yanxue, HUAI Libo\*, CUI Rongyi

( College of Engineering, Yanbian University, Yanji 133002, China )

**Abstract:** Aiming at the problem of how to use data mining technology to analyze and guide users to improve their learning behavior, this paper proposes a learning behavior analysis method based on association rules of degree of interest. Firstly, the  $K$ -means clustering method is adopted to quickly summarize the learning state of users. Secondly, strong rules between learning behavior and learning effect are obtained by association rule algorithm with degree of interest. Taking the user learning data provided by edX platform as an example, the verification results show that the number of strong rules obtained by the algorithm with degree of interest is reduced by 40.9% compared with the traditional association rule algorithm. At the same time, the method can obtain the specific relationship between learning behavior factors and learning effects, which is helpful to guide users to improve learning behavior.

**Keywords:** online course; learning behavior; clustering; association rules; interest measure

## 0 引言

随着教育信息化的推进,数字化学习已经成为当今学习者的重要学习方式.同时,数据挖掘技术的应用促进了学习分析从传统的经验性向客观性发展,为研究学习者的个性化发展提供了新的技术支持<sup>[1]</sup>.近年来,如何采用数据挖掘技术对全数据环境进行分析以获得直接、客观的教育评价

和学习分析成为学者们关注的研究热点.

教育数据挖掘是基于大量的学生个体相关数据信息的基础上,分析挖掘出隐含于这些数据背后的各类信息,使其更加具有针对性和个性化<sup>[2]</sup>.常用的教育数据挖掘方法有聚类分析、决策树、关联规则等.其中,聚类分析方法常用于学习行为特征分析<sup>[3]</sup>、判断影响成绩的因素<sup>[4]</sup>、寻找成绩评价中存在的问题<sup>[5]</sup>等.决策树算法常用于建立学生

成绩分析预测模型<sup>[6-7]</sup>. 关联规则常用于对不同学生课程的成绩进行关联分析,找出课程间的相互影响关系,为学生推荐课程或分析影响成绩的重要因素等<sup>[8-10]</sup>. 目前,相关研究大多仅用数据挖掘中的单一算法对成绩进行分析,得到的结果不够明确,难以直接用于指导改善学习行为. 例如,通过决策树可以找出影响分类的关键因素,却无法得知各项间的关联;而关联规则可得到各项间的关联,却无法说明它们之间的内在影响关系. 本文以在线课堂环境下用户的学习行为数据为研究对象,采用含兴趣度指标的关联规则算法对学习行为数据进行分析,寻找学习者的学习行为与学习效果之间的深层关系,以为学习者提供明确的学习指导.

## 1 相关算法概述

### 1.1 聚类分析

聚类是将抽象对象的集合组成为由类似的对象组成的多个类的过程. 聚类生成的类是一组数据对象的集合,聚类分析的原理是使属于同一类别的个体之间距离尽可能小,而不同类别的个体之间距离尽可能大. 目前主要的聚类算法可以划分为:划分法、层次法、基于密度的方法、基于网格的方法和基于模型的方法<sup>[11]</sup>. K-means 算法是一种典型的扁平聚类算法,是划分法中应用最为广泛的算法之一. 该算法的主要目标是最小化各元素到其簇中心的欧式距离平方的平均值,具有简单、快速的优点,可以对大型的数据集进行快速分类. 聚类准则函数用于衡量聚类结果,通常是用数据集中所有对象与各自所在簇的簇中心误差平方和来计算. 当平方误差和足够小时,即表示可以结束聚类操作. 聚类准则函数的表达式为

$$RSS = \sum_{i=1}^k \sum_{p \in c_i} |p - m_i|^2, \quad (1)$$

其中  $c_i$  表示第  $i$  类数据对象的集合,  $p$  是簇  $c_i$  中的数据对象,  $m_i$  是簇  $c_i$  的平均值,  $k$  表示该数据集可以划分为  $k$  个簇. 聚类分析可作为数据挖掘的一个模块,也可作为其他挖掘算法的预处理步骤.

### 1.2 关联规则

传统关联规则<sup>[12]</sup> 是表示项集  $X$  与项集  $Y$  的某种相关性,形如  $X \Rightarrow Y$  的蕴涵式,由支持度  $s$  和

置信度  $c$  决定. 规则  $X \Rightarrow Y$  在事务集  $D$  中成立. 支持度  $s$  是  $D$  中事务包含  $X$  和  $Y$  的百分比,即概率  $P(X \cap Y)$ ,其表达式为

$$s(X \Rightarrow Y) = P(X \cap Y). \quad (2)$$

置信度  $c$  是  $D$  中事务包含  $X$  的同时也包含  $Y$  的百分比,即条件概率  $P(Y|X)$ ,其表达式为

$$c(X \Rightarrow Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)}. \quad (3)$$

Apriori 是经典的关联规则算法之一,其包括寻找频繁项集和寻找强规则两部分. 寻找频繁项集是算法核心,包含连接、剪枝两步操作. Apriori 算法的基本思想是通过多遍扫描数据库找出全部频繁项集,从 1-项频繁集开始,递归地产生 2-项频繁集、3-项频繁集,如此下去直到产生所有的频繁项集. 最后,利用频繁项集构造出满足最小置信度的强规则.

传统关联规则算法主要考虑支持度和置信度指标,通过满足大于最小支持度和置信度来获得强关联规则,但该方法有时难以解释其规则的实际意义. 因此,学者们引入了“兴趣度”度量值,修剪无用的规则. 目前兴趣度模型主要有基于模板的兴趣度模型、基于概率相关性的兴趣度模型、基于信息量的兴趣度模型和基于差异思想的兴趣度模型等<sup>[13]</sup>,这些模型由于是从不同的角度对兴趣度进行评价,因此只适用于不同的实际问题.

基于概率相关性的兴趣度模型<sup>[14]</sup> 是从统计独立性检查的角度出发,在关联规则的置信度和支持度基础上增加一个新的相关性约束,以将不满足条件的关联规则删除.  $X$  和  $Y$  的相关性计算公式为

$$\text{Intr}(X \Rightarrow Y) = \frac{P(X \cap Y)}{P(X)P(Y)}. \quad (4)$$

式(4)中的相关性计算值作为兴趣度,其体现的是  $X$  和  $Y$  的密切程度.  $\text{Intr}(X \Rightarrow Y) = 1$ ,表示  $X$  和  $Y$  相互独立,它们之间没有相关性,此时  $P(X \cup Y) = P(X)P(Y)$ ;  $\text{Intr}(X \Rightarrow Y) > 1$ ,表示  $X$  与  $Y$  为正相关,  $X$  的出现会促进  $Y$  的出现;  $\text{Intr}(X \Rightarrow Y) < 1$ ,表示  $X$  与  $Y$  为负相关,  $X$  的出现会减少  $Y$  的出现. 在实际应用中,当关联规则的后件为单数据项时具有较为明确的决策指导意义,因此为保证规则的应用价值,在算法实现过程中只挖掘形如

$X \Rightarrow Y$  的关联规则, 这样可以减少大量的冗余关联规则, 提高算法效率.

2 基于兴趣度的学习行为分析方法

2.1 基于兴趣度的学习行为分析

传统的关联规则挖掘算法在分析学习效果的影响因素时, 通常仅考虑支持度和置信度指标<sup>[15]</sup>, 而且置信度只考虑  $X$  出现时  $Y$  的出现概率, 而未考虑  $X$  未出现时  $Y$  的出现概率, 因此在挖掘时会得到大量的冗余规则, 难以实用. 因此, 本文采用含有兴趣度指标的关联规则算法对学习行为进行分析, 以获得属性间更多的信息.

假设学生的一系列学习行为属性为集合  $A = \{A_1, A_2, \dots, A_m\}$ , 每个属性有  $k$  个不同等级的具体取值. 根据实际学习情况, 属性不同  $k$  取值不同. 假设学生的每条学习行为数据对应的学习成绩为  $Z$ , 并且  $Z$  按分数值划分为  $n$  个等级, 即  $Z = \{Z_1, Z_2, \dots, Z_n\}$ . 在分析学习行为过程中, 本文引入基于概率相关性的兴趣度模型思想, 通过计算兴趣度值分析学习行为属性与学习成绩之间的深层关系. 一般情况下, 学习行为总量为某一具体常数, 则属性间的兴趣度计算过程可由式(5)所示:

$$\text{Intr}(A_k^m \Rightarrow Z_n) = \frac{\text{count}(A_k^m \cap Z_n)}{\text{count}(A_k^m) \text{count}(Z_n)}. \quad (5)$$

其中  $\text{count}$  为统计数目,  $\text{count}(A_k^m \cup Z_n)$  为某具体行为属性与对应成绩等级共同在数据集中出现的次数,  $\text{count}(A_k^m)$  和  $\text{count}(Z_n)$  分别为该行为属性和对应成绩在数据集中各自出现的次数.  $\text{Intr}(A_k^m \Rightarrow Z_n)$  可以反映属性间的具体关系: 若其值等于 1, 则认为两者之间不存在相互影响关系; 若其值大于 1, 则认为该行为数据属性的存在会促进成绩等级的提高; 若其值小于 1, 则认为该行为数据属性的存在会抑制成绩等级的提高. 对兴趣度的计算结果进行分析, 可以解释学习行为与学习效果之间的深层关系, 进而可指导学生改善学习行为.

2.2 具体算法步骤

数据挖掘的过程一般包括 4 个部分: 数据收集、数据预处理、数据分析和结果解释. 关联规则算法是通过挖掘频繁项集来发现属性间的联系, 但若数据量大产生的规则也就越多, 用户很难观

察到某些细化区域的隐含规则, 因此本文将聚类分析作为数据挖掘的一个步骤. 首先对样本数据进行聚类将区域细化, 然后对不同簇类的数据进行关联规则挖掘, 以此提高挖掘效率.

本文采用基于兴趣度的关联规则算法进行学习行为分析的主要步骤如下:

1) 获取用户的原始学习行为数据, 并进行数据预处理, 包括数据清洗、数据集成、数据离散化等操作, 预处理后的数据存入数据库, 形成样本数据集;

2) 采用  $K$ -means 算法进行聚类, 利用公式(1)选取聚类簇数, 将数据区域细则化, 生成相互区分的类. 以学习成绩作为学习效果的依据, 对各类学习行为和学习效果进行归纳分析;

3) 采用基于兴趣度的关联规则算法对各区域数据进行挖掘, 利用式(2)和式(3)得到影响学习效果的学习行为因素, 然后根据式(5)计算结果, 分析学习行为与学习效果之间的深层联系.

3 实验结果与分析

3.1 数据预处理

实验数据来自 edX 平台提供的 MITx 的 2013 年春季编号为 8.02x 的课程学习记录, 该数据集含有学习者从注册到最后结业成绩的所有学习数据, 共计 18 579 条. 实验主要提取的学习特征分别是: 是否访问课件标签( $A$ ), 访问课程是否过半( $B$ ), 互动次数( $C$ ), 视频播放次数( $D$ ), 互动的章节数( $E$ ), 论坛发帖数( $F$ ), 是否获得证书( $G$ ), 成绩结果( $Z$ ). 为提高数据挖掘的效率, 首先进行数据预处理操作, 将原始数据离散化, 获得的部分学习特征数据如表 1 所示.

表 1 学习特征表

	A	B	C	D	E	F	G	Z
U1	1	0	1	1	1	0	0	0
U2	1	1	1	1	4	0	0	1
U3	0	0	0	0	0	0	0	0
U4	1	0	1	1	1	0	0	0
U5	1	0	1	1	1	0	0	0

表 1 中, 每一行数值代表某一名学习者的全部学习特征, 各特征项的属性见表 2.

表 2 特征值的属性

特征项	属性值	描 述
A	0,1	是否访问过课件标签,0—否,1—是
B	0,1	访问课程是否过半,0—否,1—是
C	0,1,2,3,4,5	数值越大,互动次数越多
D	0,1,2,3,4,5	数值越大,视频播放次数越多
E	0,1,2,3,4,5	数值越大,互动的章节数越多
F	0,1,2,3	数值越大,论坛发帖数越多
G	0,1	是否获得课程证书,0—否,1—是
Z	0,1,2,3,4,5	数值越大,获得的成绩级别越高

3.2 聚类分析

实验以 Eclipse 环境为平台,用 Python 作为开发语言,采用 K-means 算法对获得的学习特征进行聚类分析. 首先,通过聚类准则函数确定最佳的聚类簇数,其结果如图 1 所示.

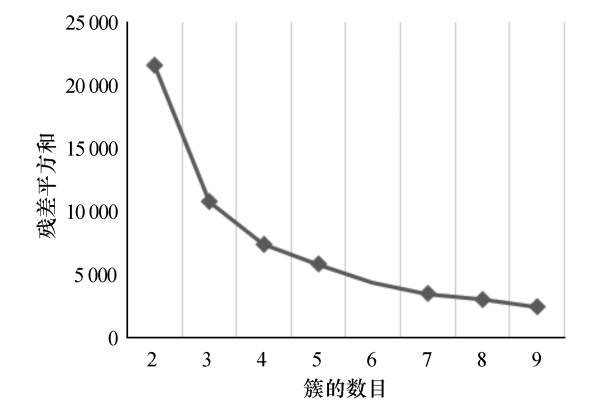


图 1 聚类的数目

由图 1 可以看出,曲线呈不断下降趋势,但结合实际情况可知聚类数不可能取无限小的值,否则失去研究意义. 当簇的数目为 3 时,曲线变化率最大,即聚类效果最好,因此本实验选取聚类数为 3. 聚类结果如表 3 所示,表中列举了每类含有的主要特征项,括号内的数值为具体人数.

从表 3 可以看出,第 1 类消极型学习者几乎

没有浏览过课件和视频等学习内容,并且几乎没有有过互动,学习质量很差,没有获得证书. 第 2 类被动型学习者虽然大多数浏览过课件和视频,以及有过互动学习经历,但大多数没能坚持学习到课程的一半,学习效果并不理想,也没能获得证书. 第 3 类主动型学习者都浏览过课件和视频,互动和发帖数较多,而且能够坚持长时间学习,因此这类学习者的学习效果较好,大多获得了相应的课程证书.

表 3 聚类结果

特征	聚类结果(人数)		
	第 1 类	第 2 类	第 3 类
访问课件	A0(4 656)	A1(12 934)	A1(851)
课程过半	B0(4 783)	B0(12 761) B1(229)	B1(828)
互动次数	C0(1 139) C1(3 644)	C1(12 945)	C1(364) C2(350) C3(113)
视频播放次数	D0(4 782)	D0(1 864) D1(11 062)	D1(697) D2(115) D3(16)
互动章节数	E0(4 690)	E1(10 968) E2(1 615) E3(332)	E4(224) E5(535)
论坛发帖数	F0(4 783)	F0(12 888) F1(48)	F0(806) F1(31) F2(8)
证书	G0(4 783)	G0(12 945)	G0(424) G1(427)
成绩	Z0(4 783)	Z0(11 368) Z1(1 547)	Z2(188) Z4(206) Z5(213)

注:第 1 类约占总人数的 25%,第 2 类约占总人数的 70%,第 3 类约占总人数的 5%.

3.3 关联规则分析

为找出影响学习效果的重要因素,分别采用传统的 Apriori 算法和含有兴趣度的改进算法对不同类型学习者的学习特征数据进行挖掘,获得的关联规则数目如表 4 所示.

表 4 不同关联规则算法的实验结果

类别	传统 Apriori 算法 关联规则数	含兴趣度的 Apriori 算法关联规则数			
		总数	负相关	相互独立	正相关
消极型	392	224	72	152	0
被动型	392	224	56	35	133
主动型	126	90	34	27	29

实验结果显示,采用含兴趣度的算法获得的强规则数目比传统 Apriori 算法减少了 40.9%. 学

习成绩作为学习效果的重要体现,分析与其相关的强规则可获知学习者的学习行为与学习效果之



间的关系. 由于大多数学习者属于被动型学习类型, 因此本文以被动型学习者为例进行分析. 被动型学习类型的部分强规则如表 5 所示.

表 5 部分强规则

序号	强规则	置信度	兴趣度
1	$A1 \Rightarrow Z0$	0.878 07	0.999 88
2	$D1 \Rightarrow Z0$	0.871 54	0.992 44
3	$G0 \Rightarrow Z0$	0.878 17	1.000 00
4	$B0 \Rightarrow Z0$	0.887 22	1.010 30
5	$E1 \Rightarrow Z0$	0.939 91	1.070 31
6	$C1 \Rightarrow Z0$	0.879 09	1.001 05
7	$F0 \Rightarrow Z0$	0.879 73	1.001 77

由表 5 中的置信度可知, 所选择的学习特征都是影响学习成绩的重要因素. 由  $G0 \Rightarrow Z0$  的兴趣度为 1.0 可知, 是否获得证书和成绩的关系是相互独立的, 不能以成绩优劣决定是否能获得证书. 学习特征  $A$ 、 $D$  与  $Z$  之间的兴趣度值均小于 1, 即访问课件、播放视频与成绩的关系为负相关, 说明当增多访问课件、播放视频等行为时, 成绩为 0 分的情况会减少; 而特征  $B$ 、 $C$ 、 $E$ 、 $F$  与  $Z$  之间的兴趣度值均大于 1, 即访问课程的次数不过半, 互动次数少、学习的章节数少、不发帖讨论等与成绩的关系为正相关, 说明这些情况的出现会增加成绩为 0 分的情况.

4 结论

本文以在线课堂的用户学习行为数据为研究对象, 通过引入兴趣度指标的关联规则算法研究了学习行为因素与学习效果之间的关系. 实验结果表明, 相比传统关联规则本文方法可有效去除冗余规则, 并且可得出规则前后件的具体联系, 有利于指导用户改善学习行为. 影响学习效果的因素较为复杂, 本文仅对在线学习用户的部分学习行为因素进行了分析, 今后将考虑网络环境、学习资源等其他因素对学习行为因素的影响, 以及提

高数据挖掘算法的准确率, 以更有效地分析学习行为因素之间的深层关系, 提高在线学习用户的学习效果.

参考文献:

[1] 刘凤娟. 大数据的教育应用研究综述[J]. 现代教育技术, 2014, 24(8): 13-19.

[2] ALGARNI A. Data mining in education[J]. International Journal of Advanced Computer Science & Applications, 2016, 7(6): 456-461.

[3] ANTONENKO P D, TOY S, NIEDERHAUSER D S. Using cluster analysis for data mining in educational technology research[J]. Educational Technology Research & Development, 2012, 60(3): 383-398.

[4] 田娜, 陈明选. 网络教学平台学生学习行为聚类分析[J]. 中国远程教育, 2015, 11: 38-41.

[5] 付希. 基于蚁群算法的聚类分析在学生成绩评价中的应用研究[D]. 成都: 西南交通大学, 2013.

[6] 董欢. 决策树技术在高校学生成绩分析中的应用研究[D]. 西安: 西安电子科技大学, 2012.

[7] 刘志妩. 基于决策树算法的学生成绩的预测分析[J]. 计算机应用与软件, 2012(11): 312-314.

[8] 杨财英. Apriori 算法及其在学生成绩分析中的应用研究[D]. 湖南: 湖南大学, 2015.

[9] 朱茜. 基于学生成绩关联分析的个性化选课推荐应用研究[D]. 武汉: 华中师范大学, 2017.

[10] ZHONG R, WANG H. Data association rules in analyzing performance level of college students [C]//International Conference on Applied Informatics & Communication. Springer, Berlin, Heidelberg, 2011, 226: 454-458.

[11] 严勇. 数据挖掘中聚类分析算法研究与应用[D]. 成都: 电子科技大学, 2007.

[12] 崔妍, 包志强. 关联规则挖掘综述[J]. 计算机应用研究, 2016, 33(2): 330-334.

[13] 张玉芳, 熊忠阳, 彭燕, 等. 基于兴趣度含正负项目的关联规则挖掘方法[J]. 电子科技大学学报, 2010, 39(3): 407-411.

[14] 丁一, 付弦. 基于兴趣度的关联规则挖掘研究[J]. 情报科学, 2011, 29(6): 939-942.

[15] 吴青, 罗儒国, 王权于. 基于关联规则的网络学习行为实证研究[J]. 现代教育技术, 2015, 25(7): 88-94.