Hans Journal of Data Mining 数据挖掘, 2024, 14(3), 189-206
Published Online July 2024 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2024.143018

Hans汉斯

关联规则分析中兴趣度量Lift与Conviction的关系探讨及教育数据验证

万　鑫*，李梓如*，李裕梅

北京工商大学数学与统计学院，北京

收稿日期：2024年6月9日；录用日期：2024年7月9日；发布日期：2024年7月18日

摘　要

关联规则分析是数据挖掘中最常用的研究方法之一。在关联关系的发现过程中兴趣度量是关联规则发现的理论基础，它可以度量规则的重要程度，其中Lift和Conviction这两个度量在数据分析中被广泛应用于筛选关联规则。本文对这两种兴趣度量进行了研究。首先，提出并证明了当后项集固定时，Conviction取值随Lift取值单调增加，且Conviction (Lift)是一个凸函数。然后，证明了当Confidence固定时，Conviction取值随Lift取值单调增加，且Conviction (Lift)是一个凹函数。最后，综合以上两个方面，得到一个重要结论：当后项集保持不变或当Confidence固定时，根据Conviction和Lift筛选出来的规则都是相同的。最后，利用某高校数学类专业三个年级的成绩数据进行了定理及相应结论的验证。

关键词

关联规则分析，Lift，Conviction，函数关系，数据分析验证

Exploration of the Relationship between Interest Measures Lift and Conviction in Association Rule Analysis and Education Data Validation

Xin Wan*, Ziru Li*, Yumei Li

School of Mathematics and Statistics, Beijing Technology and Business University, Beijing

Received: Jun. 9th, 2024; accepted: Jul. 9th, 2024; published: Jul. 18th, 2024

*共第一作者。

文章引用: 万鑫, 李梓如, 李裕梅. 关联规则分析中兴趣度量 Lift 与 Conviction 的关系探讨及教育数据验证[J]. 数据挖掘, 2024, 14(3): 189-206. DOI: 10.12677/hjdm.2024.143018

Abstract

Association rule analysis is one of the most active research methods in data mining. In the process of finding association relationships, interest measures are the theoretical basis and can measure the significance of rules, where Lift and Conviction are widely used in data analyses to find association rules. This paper studies these two measures. First, it is proven that when the Consequent is fixed, the value of Conviction increases monotonically with the value of Lift, and Conviction is a convex function of Lift. Second, when Confidence is fixed, the value of Conviction increases monotonically with the value of Lift, and Conviction is a concave function of Lift. Then, integrating the above two aspects, we obtain an important conclusion: when the Consequent remains fixed or when the Confidence value is fixed, the rules selected by Conviction are the same as those selected by Lift. Finally, the theorems and the corresponding conclusion are verified by using the achievement data of three grades of mathematics major in a university.

Keywords

Association Rule Analysis, Lift, Conviction, Functional Relationship, Data Analysis Validation

1. 引言

关联规则挖掘[1]是数据挖掘领域的一个重要研究方向，它被广泛用于在大规模数据库中寻找共现的项目，例如用于购物篮分析[2]、挖掘蛋白质-DNA 结合序列模式[3]、教育中的学生管理[4]、提取和可视化临床事件[5]、动物迁移优化[6]等等。为了更好地挖掘感兴趣的关联规则，已经提出了许多兴趣度量，例如 Support [1]，Confidence [1] [7]，Lift [8]，Conviction [2]，以及许多其他有趣的度量[9] [10]。

当前的关联规则挖掘主要在 Support-Confidence 框架下进行。然而，仅使用支持度和置信度进行挖掘通常无法满足研究人员的需求。因此，在关联规则挖掘过程中，学者们经常增加使用提升度、确信度等兴趣度量，或者构建新的兴趣度量来挖掘自己感兴趣的规则。但是，几乎所有的兴趣度量都是基于前件的支持度、后件的支持度、前件和后件共现的支持度以及对立事件的支持度来计算的[11]。随着这类研究的持续增长，兴趣指标的数量也飞速增加。因此，这些不同的兴趣度量之间是否存在相关性以及是否有必要建立如此多的兴趣度量需要进一步研究。兴趣度量的增加阿使得没有深入研究过兴趣度量的人很难在众多兴趣度中选择合适的一个来挖掘感兴趣的关联规则[9] [11]。

事实上，如果能够探索不同兴趣度量之间的关系，特别是在特定条件下一种度量是否可以被另一种度量取代，将对研究人员在关联规则挖掘中起到有效的指导作用，并帮助他们减少挖掘关联规则时需要考虑的兴趣度量，从而降低了难度。因此，本研究的目的是探索兴趣度量之间的关系，为关联规则挖掘领域提供一种更清晰可行的方法，帮助研究人员在众多兴趣度量中选择最合适的度量，使挖掘过程更高效，并使关联规则挖掘广泛应用于不同领域和场景。

在许多兴趣度量中，文献[12]-[17]认为支持度、置信度、提升度和确信度是发现关联规则最常用的兴趣度，它们被用于各种问题。同时，这四个兴趣度也被写入 R 语言和 Python 软件的关联规则分析包中，供数据分析师在选择一些参数后使用。文献[12] [18]提出了标准化的 Lift，它能够比较不同的事务集间挖

掘到的关联规则的感兴趣程度。除了发现数据中的一般关联之外，提升度和确信度也用于数据分类[9] [19] [20]。Azevedo 和 Jorge [9]在 UCI 17 数据集的分类过程中应用了多种兴趣度来选择和确定规则，并对分类结果进行了分析。Wei Song 等人[21]探讨了支持度、置信度和确信度的理论，并讨论了一种规则与另一种规则之间确信度的关系；王枭翔[16]证明了提升度与另一种兴趣度量之间的等价性；Fuguang Bao 等人[17]对支持、信心、提升和信念进行了一些理论研究，提出了双向支持(Bi-Support)、双向提升度(Bi-Lift)和双向置信度(Bi-Confidence)。

通过研究，我们发现在一定条件下，提升度和确信度在关联规则挖掘中可以替代使用。主要贡献为：1) 在置信值或后项集固定的情况下，发现了确信度和提升度之间的两个函数关系，一个是凸函数，另一个是凹函数，并且在这两个函数中，确信度都随提升度单调增加；并得出结论：在固定置信度或后项集的情况下，通过确信度筛选的关联规则与通过提升度筛选的关联规则是相同的。2) 通过教育数据和开放数据中的关联规则分析严格验证了确信度和提升度的函数关系以及相应的结论。

2. 相关定义及理解

关联规则分析里经典的算法是 Apriori 算法，使用一些规则挖掘某些事物发生的是否频繁发生，或者某件事情 X 的发生是否能够引起另外一件事情 Y 的频繁发生，这里的 X 叫前件(Antecedent)、Y 叫后件(Consequent)。在分析前件和后件关系的过程中，关联规则的兴趣度量用于衡量一条规则是否准确地显示了数据集中包含的规律，其中，支持度(Support)、置信度(Confidence)、提升度(Lift)和确信度(Conviction)是常见的度量方式[16]。

定义 1 (项集函数) [22]一个将标识符集映射到项集的映射 $i:2^{\mathcal{T}}\to 2^{\mathcal{I}}$。定义如下：

$$i(T)=\left\{x\in\mathcal{I}\middle|\forall t\in T,t\text{ 所对应的项集包含 }x\right\}.$$

其中，对于一个集合 X，2^X 表示 X 的幂集；$T\subseteq\mathcal{T}$，且 $i(T)$ 是事务标识符集 T 中所有事务的公共项的集合。

定义 2 (标识符集函数) [22]一个将项集映射到标识符集的映射 $t:2^{\mathcal{I}}\to 2^{\mathcal{T}}$。定义如下：

$$t(X)=\left\{t\middle|t\in\mathcal{T},t\text{ 所对应的项集包含 }X\right\}.$$

定义 3 (支持度) [22]一个项集 X 的支持度为：

$$Support(X)=\frac{\left|\left\{t\middle|\langle t,i(t)\rangle\in D\wedge X\subseteq i(t)\right\}\right|}{|D|}=\frac{|t(X)|}{|D|}=P(X),$$

$$Support(X\to Y)=\frac{|t(XY)|}{|D|}=P(XY),$$

其中，$|D|$ 表示 D 中事务个数。

这个定义中其实发生了 X 的定义转换，$Support(X)$ 与 $P(X)$ 中的 X 其实一个是项集另一个是随机事件。如果 $P(X)$ 中的 X 用 X' 表示，那么定义应该按照以下方式进行书写：

假设随机事件 X' 表示“项集 X 中的所有元素共同出现”，那么 $Support(X)=P(X')$。为了便于书写将 X' 与 X 全部书写为 X。

假设 $X=\{a,b,c\}$，那么 $Support(X)=P(X)=P(a\text{ 出现},b\text{ 出现},c\text{ 出现})$，也就是 X 的支持度是包含 X 中的每个项出现的联合概率 $P(abc)$。

定义 4 (置信度) [22] X 发生的前提下，Y 发生的概率称为置信度：

$$Confidence(X \to Y) = P(Y \mid X) = \frac{P(XY)}{P(X)} = \frac{Support(X \to Y)}{Support(X)}. \tag{1}$$

定义 5 (提升度) [16] X 出现的前提下 Y 的出现的概率与数据库中 Y 出现的概率的比值，或 X 和 Y 共同出现的概率与 X 和 Y 分别出现的概率乘积的比值称为提升度：

$$Lift(X \to Y) = \frac{Support(X \to Y)}{Support(X)Support(Y)} = \frac{Confidence(X \to Y)}{Support(Y)} = \frac{P(XY)}{P(X)P(Y)}. \tag{2}$$

定义 6 (确信度) [2]数据库中 Y 不出现的概率与 X 出现的前提下 Y 不出现概率的比值称为确信度：

$$Conviction(X \to Y) = \frac{1 - Support(Y)}{1 - Confidence(X \to Y)} = \frac{P(\overline{Y})}{P(\overline{Y} \mid X)} = \frac{1 - P(Y)}{1 - \frac{P(XY)}{P(X)}}. \tag{3}$$

理解 1：Lift 是一种双向关系，很多文献里关于它的定义都型如定义(5)那样，实际上，还可以写成如下这种形式：

$$Lift(X \to Y) = \frac{P(XY)}{P(X)P(Y)} = Lift(Y \to X)，或者：\ Lift(X \leftrightarrow Y) = \frac{P(XY)}{P(X)P(Y)}。$$

这是因为提升度不但表达了 $X \to Y$，也表达了 $Y \to X$，是 X 和 Y 之间双向关系的展现。

理解 2：这里的 Conviction 有人叫它“出错率”或者“错误率”[23] [24]，我们认为不合适，因为这个指标的取值，不是越小越好(错误率的直观意思是：值越小越好)，后面可以看到，这个 Conviction ($X \to Y$)的取值的范围是(0, ∞)，不一定是越小越好。在文献[2]里，当 Conviction 被提出来的时候，作者针对一份人口分析数据，探讨了其取值对应着不同的有趣的规则，其中有 23,712 条规则的 Conviction 取值是大于 1.25 的，并且其中的 6732 条规则的 Conviction 取值是∞，而且他们总结出，针对他们分析的那份数据，Conviction 取值在[1.01, 5]的规则是最有趣的，最值得被关注。

理解 3：Lift 实际上，就是看 X 里 Y 发生的概率，以及整体里 Y 发生的概率，Lift 就是这两个概率的比，

$$Lift(X \to Y) = \frac{P(XY)}{P(X)P(Y)} = \frac{P(XY)/P(X)}{P(Y)}$$

如果前面概率值大于后面概率值，Lift 就大于 1，说明 X 的发生、提升了 Y 的发生；如果前面概率值小于后面概率值，说明 X 的发生没有提升 Y 的发生、或者反方向提升了、实际是降低了 Y 的发生；如果两个概率值相等，则说明 X 的发生对 Y 的发生既没有正方向提升、也没有反方向提升。文献[25]里，把 $\frac{P(XY)}{P(X)}$ 看成一个增量，这个增量就能体现 X 的出现对 Y 的出现提升了多少。

理解 4：Conviction 确信度里“确信”的意思。根据公式(3)，分子 1-$P(Y)$表示 Y 不在总体里发生的概率，分母 1 − $P(XY)/P(X)$表示 Y 不在 X 里发生的概率，如果前者概率值大于后面概率值，说明 Y 在整体里不发生的情况多于 Y 在 X 里不发生的情况，相当于 X 的出现“确信”了 Y 的出现；如果前面概率值小于后面概率值，就说明 Y 在总体里不发生的情况少于 Y 在 X 里不发生的情况，相当于 X 的出现“确信”了 Y 的不出现，即反向“确信”了 Y 的出现。

3. 提升度与确信度的函数关系

本节针对如何在错综复杂的兴趣度量中选取适用的兴趣度量进行关联规则挖掘这一问题进行深入探

讨。在固定后项和固定置信度这两种不同条件的实际情况下，给出了关于提升度和确信度间的关系。

定理 1 设后项(Consequent)不变，即 $P(Y)=c$，c 是常数，则确信度随提升度单调增加，且 Conviction (Lift)是一个凸函数。

证明：

将 $Conviction(X\to Y)=\dfrac{1-P(Y)}{1-\dfrac{P(X,Y)}{P(X)}}$ 的分子分母同时乘以 $P(X)$，得到：

$$Conviction(X\to Y)=\frac{P(X)-P(X)P(Y)}{P(X)-P(XY)}. \tag{4}$$

将 Lift 的定义公式(2)进行变形，可以得到：

$$P(XY)=Lift(X\to Y)P(X)P(Y). \tag{5}$$

然后将公式(5)代入公式(4)中即可得到：

$$Conviction(X\to Y)=\frac{P(X)\left[1-P(Y)\right]}{P(X)\left[1-Lift(X\to Y)P(Y)\right]}. \tag{6}$$

又因为 $P(X)\neq 0$，因此可以将公式(6)化简称为：

$$Conviction(X\to Y)=\frac{1-P(Y)}{1-Lift(X\to Y)P(Y)}. \tag{7}$$

公式(7)是确信度和提升度的函数表达式。由定理假设 $P(Y)=c$，并且将确信度和提升度赋予数学符号为：$Conviction(X\to Y)=g(t)$，$Lift(X\to Y)=t$，因此公式(7)可以被表示为：$g(t)=\dfrac{1-c}{1-ct}$。

首先确定 t 的取值范围，根据提升度的取值范围可知 $0\le t\le \min\left\{\dfrac{1}{P(X)},\dfrac{1}{P(Y)}\right\}$，又因为存在间断点：$t=\dfrac{1}{P(Y)}$，所以 t 的取值范围是：$0\le t\le \dfrac{1}{P(X)}<\dfrac{1}{P(Y)}$ 或者 $0\le t<\dfrac{1}{P(Y)}$。这两个区间段都处于间断点的左侧，因此在后续求单调性与凹凸性的时候只考虑间断点左侧的区间。

然后通过求取 $g(t)$ 的一阶导数和二阶导数，来判断单调性与凹凸性：

$g(t)$ 的一阶导数为：

$$g'(t)=(1-c)\left(-\frac{1}{(1-ct)^2}\right)(-c).$$

因为 $0<c<1$，所以 $(1-c)>0$，$-\dfrac{1}{(1-ct)^2}<0$，因此 $g'(t)>0$，即 $g(t)$ 为 t 的单调增函数。

$g(t)$ 的二阶导数为：

$$g''(t)=\frac{2c^2(1-c)}{(1-ct)^3}=\frac{2c^2(1-c)}{\left(1-\dfrac{P(XY)}{P(X)}\right)^3}=\frac{2c^2(1-c)}{\left(1-Confidence(X\to Y)\right)^3}.$$

因为 $c<1$，并且 $Confidence(X\to Y)<1$，所以 $g''(t)>0$，即 $g(t)$ 是 t 的凸函数。

综上所述，当后项不变时，确信度是提升度的单调增函数，并且函数 Conviction (Lift)是一个凸函数。

定理证明完毕。

定理 2 设置信度是不变的，即 $Confidence(X \to Y) = c$，c 是常数，则确信度随提升度单调增加，且 Conviction (Lift)是一个凹函数。

证明：

将 Lift 的定义公式(2)进行变形：

$$P(Y) = \frac{P(XY)}{Lift(X \to Y)P(X)}. \tag{8}$$

然后将公式(8)代入到确信度定义(3)中，得到：

$$Conviction(X \to Y) = \frac{1 - \frac{P(X,Y)}{P(X)Lift(X \to Y)}}{1 - \frac{P(X,Y)}{P(X)}}. \tag{9}$$

将置信度定义(1)代入到公式(9)中可知：

$$Conviction(X \to Y) = \frac{1 - \frac{Confidence(X \to Y)}{Lift(X \to Y)}}{1 - Confidence(X \to Y)}, \tag{10}$$

整理公式(10)得到确信度和提升度的函数关系式：

$$\begin{aligned} &Conviction(X \to Y) \\ &= \frac{1}{1 - Confidence(X \to Y)} - \frac{Confidence(X \to Y)}{1 - Confidence(X \to Y)} \cdot \frac{1}{Lift(X \to Y)} \end{aligned} \tag{11}$$

由定理假设 $Confidence(X \to Y) = c$，并且将确信度和提升度赋予数学符号为：$Conviction(X \to Y) = h(t)$，$Lift(X \to Y) = t$。因此公式(11)可以被表示为：$h(t) = \frac{1}{1-c} - \frac{c}{1-c} \cdot \frac{1}{t}$。

$h(t)$ 的一阶和二阶导数分别为：$h'(t) = \frac{c}{1-c}\frac{1}{t^2}$，$h''(t) = -\frac{c}{1-c}\frac{2}{t^3}$。由于 $0 < c < 1$，$t > 0$，因此 $h'(t) > 0$，$h''(t) < 0$，即 $h(t)$ 为 t 的单调增函数，且是凹函数。

综上所述，当置信度保持不变时，确信度是提升度的单调增函数，并且函数 Conviction (Lift)是一个凹函数。定理证明完毕。

通过上面定理 1 和定理 2，得到结论：当 Consequent 不变或者 Confidence 不变时，用 Lift 或者 Conviction 的值排序筛选出来的规则是相同的。

得到上面的结论，是因为：根据固定的 Consequent 或者 Confidence，一共得到很多条规则，然后把规则根据对应 Lift 的从小到大排序得到顺序 1，也把规则根据对应 Conviction 从小到大排序到大顺序 2，这时候，发现顺序 1 和顺序 2 对应的规则完全一致。那么当按照大于或等于某个 Lift 值取出来规则时，也就是按照对应的 Conviction 大于或等于另外某个值取出来的规则，所以完全对应。实际上，因为 Conviction 是 Lift 的函数，那么当 Lift 取某个值时，根据函数关系，也能计算出 Conviction 的一个值，这两个值对应到同一条规则。

4. 定理的实验验证

本节利用一个私有数据集挖掘关联规则、计算相关兴趣度量取值，验证定理 1 和定理 2。通过实验

挖掘到的关联规则计算它们的兴趣度量取值，然后绘制兴趣度量间的折线图验证相关定理。

实验中所使用的私有数据是某大学 2016~2018 级数学系本科生课程成绩，在后续书写过程中称作“教育数据”。该数据选取了某大学 2016~2018 三个年级数学大类中“应用统计学”和“信息与计算科学”两个专业共 233 名学生的所有成绩，该数据的每一行是一个学生的某门课程的一次考试的各种信息，包括：学号、姓名、学制、开课学期、上课院系、班级名称、课程编号、课程名称、总成绩、原始总成绩、成绩标志、课程性质、课程属性、通选课类别、学时、学分、开课单位、考试性质、补重学期，共 19 项。在本文中主要使用学号、课程名称、总成绩、学分这四列，部分数据如表 1 所示。

Table 1. Education data display before preprocessing
表 1. 预处理前教育数据展示

学号	课程名称	总成绩	学分
122****07	大学英语(四)	0	3
122****07	剑桥商务英语(中级)	83	3
122****07	微积分(上)	64	3
122****07	管理学	0	3
122****07	英语口语(一)	83	2
122****07	经济法	80	3
122****07	思想道德修养与法律基础	41	3
122****07	英语口语(二)	74	2
122****07	计算机技术	60	3
122****07	会计学原理(全英)	88	3
…	…	…	…

考虑到原始数据存在：隐私信息未加密、数据格式不匹配、同一门课程多次考试(挂科)、学生在毕业前退学或转专业、同一门课程多学期连续开课等情况，需要对数据进行成绩预处理和基于处理后数据构建独热编码 2 种操作。

4.1. 数据预处理

成绩预处理操作如流程图 1 所示，共分为 8 个步骤。前 7 个步骤是将三个年级的数据按照相同的过程分开进行操作，第 8 个步骤将所经过处理的三个年级成绩进行合并。

步骤 1：计算考试次数。

步骤 2：计算成绩权重与核查考试总成绩。

步骤 2-1：计算成绩权重。

根据每个学生在某一门课程中的考试次数来确定权重矩阵 $\boldsymbol{\omega}$。权重的计算公式为 $\omega_{ij}=1-\eta\left(n_{ij}-1\right)$，其中 ω_{ij} 是学生 i 在课程 j 的权重；η 是惩罚参数；n_{ij} 是学生 i 的课程 j 的考试次数。在这篇文章中，将惩罚参数 η 的取值设置为 0.03。因为每个年级的学生人数不一样，以及每个年级学生选课情况有差异，所以，i, j 的取值范围在每个年级中是不同的，具体如表 2 所示。

这里，η 是一个惩罚参数，如果通过多次考试获得 60 分，它用于降低分数，因为，在同一课程中，通过正常期末考试获得 60 分的学生比通过多次考试得到 60 分的学生对本门课程中的掌握程度更高，为了区别这两种情况下学生对知识的掌握程度，将后者的 60 分进行惩罚降低到 60 分以下。在正常的期末

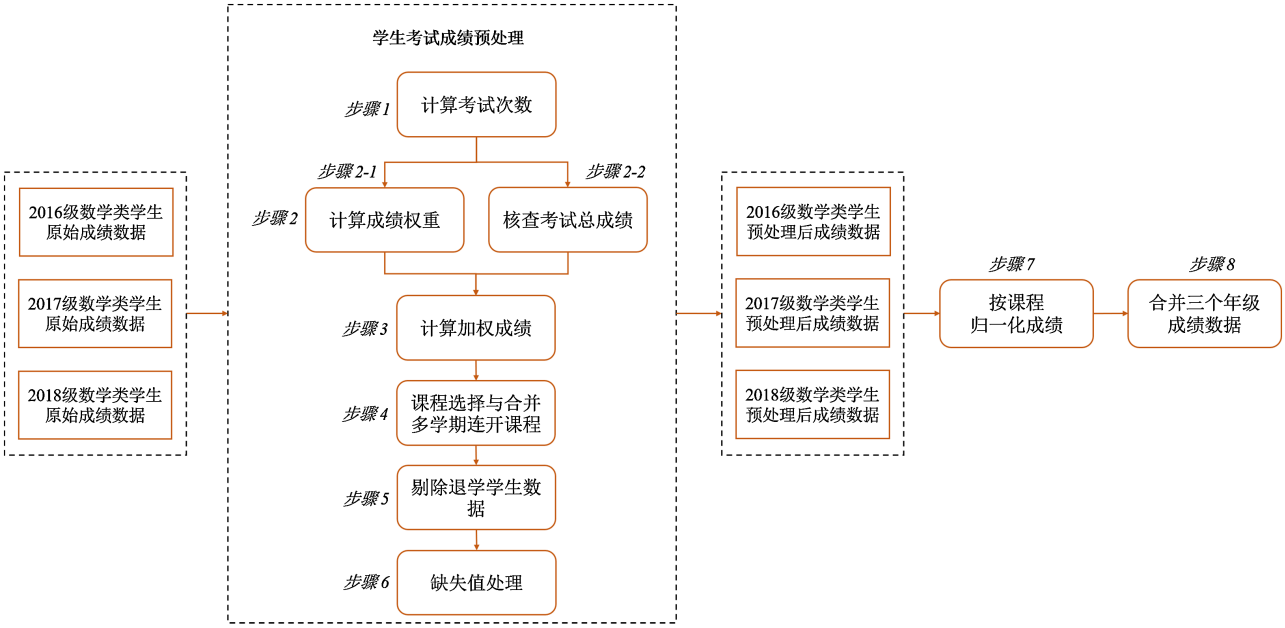

Figure 1. Educational data preprocessing flowchart
图 1. 教育数据预处理流程图

Table 2. The value range of parameters *i*, *j*
表 2. 参数 *i*，*j* 取值范围

参数	2016 级	2017 级	2018 级
i	1, 2, 3, …, 94	1, 2, 3, …, 77	1, 2, 3, …, 62
j	1, 2, 3, …, 191	1, 2, 3, …, 197	1, 2, 3, …, 168

考试中，学生的分数几乎都分布在 40~95 分之间，但是，挂科的学生经过多次补考后，他们对知识的掌握程度得到了一定的提高，所以课程分数的下限应该有所提高，在 50~95 分更合适。为了选择合适的惩罚参数，使分数可以在 50~95 之间分布，我们通过设置 $\eta = 0.02$、0.03、0.04 进行了实验，发现，当 $\eta = 0.03$ 时几乎所有的分数都分布在 50~95，最终选取惩罚参数 $\eta = 0.03$。

步骤 2-2：核查考试总成绩。

由于这所大学补考成绩规定：① 如果参加某门课程的补考并通过考试的学生将获得最终得分为 60 分；② 如果一名学生参加了补考，但未能通过这门课程的考试，则他在这门课上的最终成绩应该是他多次考试的多个成绩中的最大值。此外，如果某个学生在某门课程中的考试次数为 0，我们认为他没有选择该课程，分数将设置为 0。

根据以上规则将每个学生的考试成绩进行核查，根据学生 i 在课程 j 的考试成绩 $scores_{ij}$，构建每个年级的成绩矩阵 **DataOld**，它的每一个分量的计算公式为：

$$DataOld_{ij} = \begin{cases} scores_{ij}, \text{考试次数} = 1 \\ 0, \text{考试次数} = 0 \\ 60, \text{考试次数} > 1 \text{并且通过考试} \\ \max\left\{scores_{ij}\right\}, \text{考试次数} > 1 \text{但没通过考试} \end{cases},$$

其中 i，j 的取值范围如表 2 所示。

步骤3：计算加权成绩。

将权重矩阵与成绩矩阵的对应位置进行乘法运算，得到加权成绩矩阵 **Data**。

步骤4：选择课程与合并多学期连开课程。

由于后续实验中只使用必修课和专业课进行关联规则分析，因此需要对数据中所需的课程进行筛选。

根据这所大学 2017 级数学大类中的“应用统计学”和“信息与计算科学”专业的培养计划，以及培养计划中对于专业必修课、公共必修课和专业选修课的分类，确定了需要进行分析的课程，如表 3 所示。

Table 3. Selected courses classification
表 3. 所选课程分类

专业必修课	公共必修课	专业选修课
《高等代数(上)》	《安全素养》	《数学实验》
《高等代数(下)》	《大学生心理健康教育》	《运筹学》
《空间解析几何》	《大学英语(一)》	《抽样调查》
《数学分析(I)》	《大学英语(二)》	《回归分析》
《数学分析(II)》	《大学英语(三)》	《数学建模》
《数学分析(III)》	《大学英语(四)》	《统计学方法及应用》
《常微分方程》	《军事理论》	《应用随机过程》
《离散数学》	《军训》	
《数值分析》	《身体素质基础训练》	
《程序语言设计》	《职业生涯规划》	
《算法与数据结构》	《思想政治理论课社会实践》	
《复变函数论》	《体育(一)》	
《概率论》	《体育(二)》	
《数理统计》	《体育(三)》	
《计算机技术》	《大学物理(上)》	
	《大学物理(下)》	
	《大学生就业指导》	
	《思想道德修养与法律基础》	
	《中国近现代史纲要》	
	《马克思主义基本原理》	
	《毛泽东思想和中国特色社会主义理论体系概论》	

表 3 中共罗列了 43 门课程，涵盖了“应用统计学”和“信息与计算科学”这两个专业在大学期间需要学习的必修课与专业选修课。同时根据表 3 可以观察到存在“高等代数”这门课程分为 2 个学期进行开设，“数学分析”分为 3 个学期开设，“大学英语”分为 4 个学期开设等情况的发生。这是由于该所学校规定一门课程可以在多个连续的学期中开设以保证教学效果，并且在多个学期中获得相应的分数。所以，在不同的学期中，它们被视为不同的课程，从而产生多个加权分数。因此，需要将连续几个学期开设的同一门课程合并为一门课程，本文采取了将同一学生在连续多个学期开设的课程的加权分数进行平均作为该课程的加权分数。这一步骤中进行合并的课程名称以及合并后的课程名称如表 4 所示。

Table 4. Details of the merged courses
表 4. 合并课程详情

合并前课程名称	合并后课程名称
高等代数(上)，高等代数(下)	高等代数
数学分析(I)，数学分析(II)，数学分析(III)	数学分析
概率论，数理统计	概率论与数理统计
大学物理(上)，大学物理(下)	大学物理
大学英语(一)，大学英语(二)，大学英语(三)，大学英语(四)	大学英语

经过这一步骤极大程度避免了如“高等代数(上)→高等代数(下)”这种冗余的规则出现，同时将无用数据删除，降低了挖掘关联规则过程中的内存占用与计算量。最终，原始选择的 43 门课程变为 35 门，用于后续关联规则挖掘。

步骤 5：剔除退学学生数据。

如果某学生的加权分数不等于 0 的课程个数小于 50 个，则该生被认为是在毕业前辍学的，因此他的所有数据被剔除。

这里，加权分数不等于 0 的课程个数小于 50 个，则相应的学生被认为退学的原因如下：“应用统计学”专业的总学分为 146，如果课程的平均学分为 3，由于 146/3 = 48.66，这意味着每个学生必须学习 49 门课程。但是，考虑到有些课程的学分是 1，少数课程的学分为 5 或 4，而且学分为 1 的课程比学分为 4 或 5 的课程多，需要学习的课程总数会增加，因此估计 50 门是合适的。在这里，我们已经根据 2016 级的数据验证了这种情况，其中有 7 名学生的加权分数不等于 0 的课程个数少于 50 个，并且这 7 名学生确实辍学了。

在剔除一些学生后，所研究的数据中还剩下 219 名学生，其中 2016 级 87 个、2017 级 72 个、2018 级 60 个。

步骤 6：处理缺失值。

有些学生可能没有选择某一门专业选修课，则这门课程的分数使用其专业必修课和公共必修课加权分数的平均值来填充。

步骤 7：按列归一化课程成绩。

为了消除不同老师给分区间不同的影响，对每个年级的成绩按照课程为一组数据进行最大最小归一化。详细计算式为：

$$\mathbf{Data}_{\cdot j}=\frac{\mathbf{Data}_{\cdot j}-\min\left\{\mathbf{Data}_{\cdot j}\right\}}{\max\left\{\mathbf{Data}_{\cdot j}\right\}-\min\left\{\mathbf{Data}_{\cdot j}\right\}}.$$

其中，$\mathbf{Data}_{\cdot j}$ 是加权成绩的第 j 列，$j=1,2,3,\cdots,35$。

步骤 8：合并三个年级的成绩。

将三个年级的归一化数据以课程为索引进行合并。得到三个年级 219 名学生 35 门课程的最终考试成绩 **SCORES**，部分数据如表 5 所示。

表 5 的每一行为每一个学生的学号以及经过选择的课程成绩，其中课程成绩是经过最大最小归一化给出的，而不是传统的百分制。具体数据处理操作在步骤 7 中进行了描述。

4.2. 基于最终考试成绩 SCORES 构建独热编码

由于用作关联规则挖掘的数据，应该是离散型 TRUE-FALSE 矩阵，即独热编码类型，需要将连续型

Table 5. The final scores recorded in variable **SCORES**
表 5. 最终考试成绩 **SCORES**

学号	概率论与数理统计	安全素养	数值分析	…	数学建模	统计学方法及应用	应用随机过程
160***223	0.65	0.71	0.62	…	0.71	0.63	0.90
160***123	0.24	0.49	0.26	…	0.74	0.55	0.25
160***328	0.41	0.29	0.74	…	0.81	0.63	0.55
160***106	0.05	0.53	0.25	…	0.05	0.04	0.05
…	…	…	…	…	…	…	…
18*****129	0.88	0.33	1.00	…	0.87	1.00	0.92
18*****221	0.56	0.22	0.41	…	0.29	0.65	0.65

教育数据进行离散化。本文通过 **SCORES** 构建了独热编码矩阵 $\mathbf{DataOH}=\left(d_{ij}\right)$，$i=1,2,\cdots,219$；$j=1,2,\cdots,35$，这里 i 代表学生，j 代表课程。矩阵 **DataOH** 中的每一个元素 d_{ij} 的通过公式(12)定义。

$$d_{ij}=\begin{cases}1, & x_{ij}\geq \mu_{\cdot j}+\sigma_{\cdot j}\\ 0, & \text{others}\end{cases}. \tag{12}$$

其中，x_{ij} 代表学生 i 在课程 j 所得的成绩，$\mu_{\cdot j}$ 代表所有学生课程 j 考试成绩的平均值，$\sigma_{\cdot j}$ 代表所有学生课程 j 考试成绩的标准差。

在本文中，将“好成绩”的标准定义为考试成绩高于“平均值 + 标准差”，是因为，假定成绩近似服从高斯分布，那么分数高于“平均值 + 标准差”的学生大概是课程中排名前 15%的学生。

4.3. 基于教育数据的实验验证

本节将利用前面预处理过的教育数据对我们前面提出来的定理和相关结论进行实验验证。同时，通过我们呈现的验证过程，也可以让其他研究人员更进一步了解和使用我们的数据集。

4.3.1. 基于教育数据的关联规则挖掘

根据数据预处理的结果 **DataOH**，初步进行关联规则挖掘。其中，我们要挖掘的是“大一大二的某些课程学习成绩好的情况下，大三的另外一些课程学习成绩也好”这样的规则。首先利用 Apriori 算法选择最小支持度为 0.06，挖掘频繁项集。然后，设定最小置信度为 0.5 进行关联规则挖掘。但是，这些挖掘的关联规则，它们的前项可能包含大三所上的课，后项也可能包含大一大二上的课。这样的规则不是我们想要的，因为它不符合学习这门课程的时间顺序。因此需要对挖掘的关联规则进行筛选，删除上述所说的规则。最后只保留了 1067 条关联规则，这 1067 条规则将用于后续基于教育数据的定理验证。其中前 5 条和后 5 条规则如表 6 所示。

表 6 是按照置信度将所挖掘到的关联规则进行降序排序后得到的。第一条关联规则是：算法与数据结构、概率论与数理统计、复变函数论→数学实验。这意味着如果某位学生的“算法与数据结构、概率论与数理统计和复变函数论”这三门课程同时学的比较好的话，那么这位学生“数学实验”这门课程也有很大的可能学的好，这也刚好符合我们平时对于这几门课程的固有认识。但是存在一些挖掘到的关联规则与我们平时对学科的认知是非常不一样的，甚至是存在一些偏差。这是由于在实验参数的设置中，为了更好的验证第三章中所提出的定理，在挖掘关联规则的过程中一些参数设置的比较小，以便得到更多的关联规则，这是以得到的关联规则质量为代价。所以会出现一些没有价值的关联规则，但这不影响后续的定理验证。

Table 6. Association rules based on educational data mining
表 6. 基于教育数据挖掘的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
算法与数据结构，概率论与数理统计，复变函数论	数学实验	0.0731	0.9412	5.0273	13.8174	0.8947
算法与数据结构，毛泽东思想和中国特色社会主义理论体系概论	回归分析	0.0685	0.9375	5.1328	13.0776	0.8889
算法与数据结构，概率论与数理统计，大学物理，复变函数论	数学实验	0.0685	0.9375	5.0076	13.0046	0.8889
…						
空间解析几何	数学实验	0.1096	0.5000	2.6707	1.6256	0.5000
大学物理	数学实验	0.1187	0.5000	2.6707	1.6256	0.5000
程序设计语言	统计学方法及应用	0.1142	0.5000	2.6707	1.6256	0.5000
数值分析，常微分方程	数学实验	0.0776	0.5000	2.6707	1.6256	0.5000

4.3.2. 兴趣度量间的单调性验证

定理 1 和定理 2 指出了不同兴趣度量间的单调关系，这一部分将以教育数据作为实验数据，根据不同定理中的假设，分别设置两种不同的假设条件，绘制定理 1 和定理 2 中所给出的不同兴趣度量间的函数图像，分析单调性和相关关系以验证我们所提定理的准确性。

按照定理 1 的假定条件，分别选择三种后项，即分别固定三个 $P(Y)$ 为常数。在实验中所选择的三个后项为“回归分析”“数学实验”“应用随机过程”。从表 6 中将满足这三种后项的关联规则分别筛选出来，如表 7~9 所示。

Table 7. Association rules based on educational data with “Regression Analysis” as the consequents
表 7. 基于教育数据后项为“回归分析”的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
算法与数据结构，毛泽东思想和中国特色社会主义理论体系概论	回归分析	0.0685	0.9375	5.1328	13.0776	0.8889
概率论与数理统计，常微分方程，毛泽东思想和中国特色社会主义理论体系概论	回归分析	0.0685	0.9375	5.1328	13.0776	0.8889
空间解析几何，概率论与数理统计，毛泽东思想和中国特色社会主义理论体系概论	回归分析	0.0639	0.9333	5.1100	12.2603	0.8824
…						
复变函数论	回归分析	0.0959	0.5122	2.8043	1.6756	0.5116
离散数学	回归分析	0.1050	0.5000	2.7375	1.6347	0.5000
马克思主义基本原理	回归分析	0.0685	0.5000	2.7375	1.6347	0.5000

Table 8. Association rules based on educational data with “Mathematical Experiment” as the consequents
表 8. 基于教育数据后项为“数学实验”的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
算法与数据结构，概率论与数理统计，复变函数论	数学实验	0.0731	0.9412	5.0273	13.8174	0.8947
算法与数据结构，常微分方程，复变函数论，概率论与数理统计	数学实验	0.0685	0.9375	5.0076	13.0046	0.8889

续表

		…				
空间解析几何	数学实验	0.1096	0.5000	2.6707	1.6256	0.5000
大学物理	数学实验	0.1187	0.5000	2.6707	1.6256	0.5000
数值分析，常微分方程	数学实验	0.0776	0.5000	2.6707	1.6256	0.5000

Table 9. Association rules based on educational data with “Applying Random Processes” as the consequents
表 9. 基于教育数据后项为“应用随机过程”的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
数值分析，数学分析，算法与数据结构	应用随机过程	0.0731	0.8421	5.1228	5.2922	0.8095
算法与数据结构，常微分方程，数学分析，数值分析	应用随机过程	0.0731	0.8421	5.1228	5.2922	0.8095
数值分析，常微分方程，离散数学，数学分析	应用随机过程	0.0685	0.8333	5.0694	5.0137	0.8000
	…					
常微分方程，复变函数论	应用随机过程	0.0731	0.5161	3.1398	1.7269	0.5152
常微分方程	应用随机过程	0.1142	0.5000	3.0417	1.6712	0.5000
概率论与数理统计	应用随机过程	0.0913	0.5000	3.0417	1.6712	0.5000

按照定理 2 的假定条件，分别选择三种置信度，即分别固定三个 $Confidence(X \to Y)$ 为常数。在实验中所选择的三个置信度为 0.5，0.5357，0.6087。从表 6 中将满足这三种置信度的关联规则分别筛选出来，如表 10~12 所示。

Table 10. Association rules based on educational data with the confidence level of 0.5
表 10. 基于教育数据置信度为 0.5 的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
数值分析，常微分方程，数学分析	抽样调查，应用随机过程	0.0639	0.5000	6.4412	1.8447	0.5000
大学物理，常微分方程，数学分析	回归分析，应用随机过程	0.0639	0.5000	5.4750	1.8174	0.5000
数值分析，常微分方程，大学物理	回归分析，应用随机过程	0.0639	0.5000	5.4750	1.8174	0.5000
	…					
程序设计语言	统计学方法及应用	0.1142	0.5000	2.6707	1.6256	0.5000
数值分析，常微分方程	数学实验	0.0776	0.5000	2.6707	1.6256	0.5000
数值分析，常微分方程	统计学方法及应用	0.0776	0.5000	2.6707	1.6256	0.5000

Table 11. Association rules based on educational data with the confidence level of 0.5357
表 11. 基于教育数据置信度为 0.5357 的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
概率论与数理统计，数学分析	回归分析，应用随机过程	0.0685	0.5357	5.8661	1.9571	0.5333
数值分析，概率论与数理统计	回归分析，应用随机过程	0.0685	0.5357	5.8661	1.9571	0.5333
数值分析，常微分方程，数学分析	回归分析，应用随机过程	0.0685	0.5357	5.8661	1.9571	0.5333
	…					

续表

数值分析，概率论与数理统计	统计学方法及应用	0.0685	0.5357	2.8615	1.7506	0.5333
数值分析，常微分方程，大学物理	数学实验	0.0685	0.5357	2.8615	1.7506	0.5333
大学物理，常微分方程，数值分析	统计学方法及应用	0.0685	0.5357	2.8615	1.7506	0.5333

Table 12. Association rules based on educational data with the confidence level of 0.6087
表 12. 基于教育数据置信度为 0.6087 的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
数值分析，常微分方程，大学物理，概率论与数理统计	回归分析，应用随机过程	0.0639	0.6087	6.6652	2.3222	0.6000
概率论与数理统计，高等代数	数学实验，运筹学	0.0639	0.6087	6.3478	2.3105	0.6000
复变函数论，高等代数	数学实验，运筹学	0.0639	0.6087	6.3478	2.3105	0.6000
	…					
大学物理，计算机技术	数学实验	0.0639	0.6087	3.2513	2.0771	0.6000
大学物理，空间解析几何，常微分方程	数学实验	0.0639	0.6087	3.2513	2.0771	0.6000
数值分析，常微分方程，大学物理，概率论与数理统计	数学实验	0.0639	0.6087	3.2513	2.0771	0.6000

根据上述描述能够将表 7~12 这 6 个表格按照三个为一组的方式进行分组，绘制不同兴趣度量间的折线图。首先，以提升度作为横轴，以确信度作为纵轴，分别将表 7~9 中的 3 组数据绘制成折线图，如图 2 所示。然后，以提升度作为横轴，以确信度作为纵轴，将表 10~12 中的 3 组数据绘制成折线图，如图 3 所示。

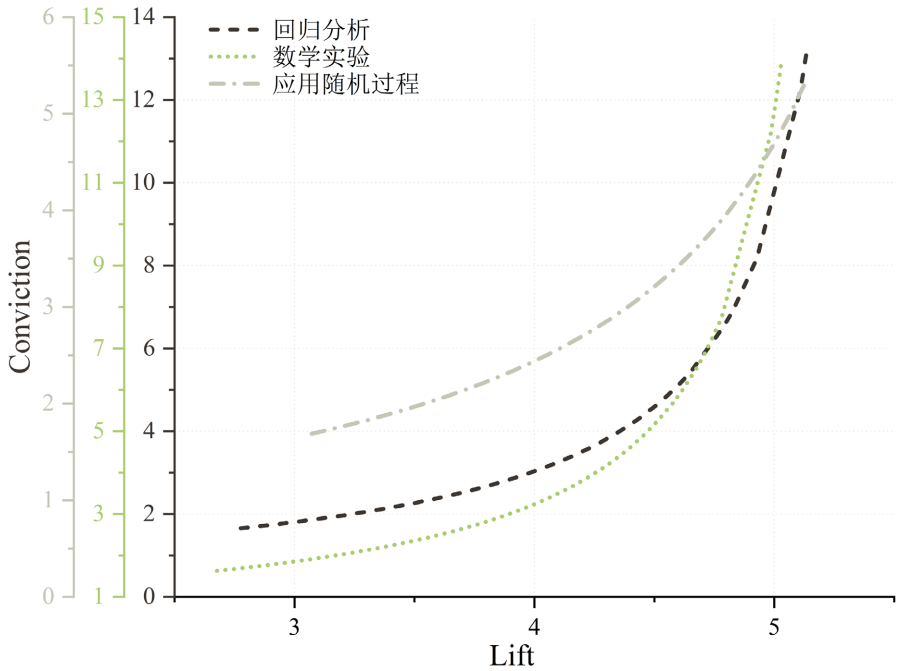

Figure 2. Lines chart of conviction and lift under different posterior terms
图 2. 不同后项下确信度与提升度折线图

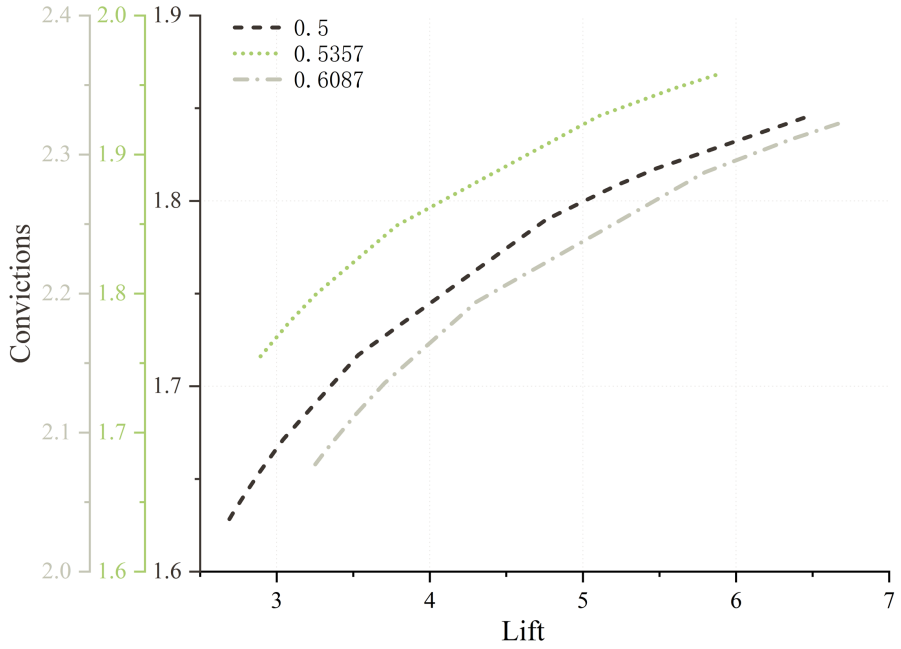

Figure 3. Lines chart of conviction and lift under different posterior terms
图 3. 不同后项下确信度与提升度折线图

从图 2 和图 3 这两幅图片中可以发现，所有折线均呈现出单调递增的趋势，符合定理 1 和定理 2 中关于确信度和提升度间单调性的描述。图 2 中确信度和提升度的函数图像呈现出凸函数的特点，因此可以认为定理 1 中给出的结论正确。图 3 中确信度和提升度的函数图像呈现出凹函数的特点，因此可以认为定理 2 中给出的结论正确。验证了两个定理阐述的确信度与提升度所构成单调递增关系。

5. 结论的实验验证

从表 6 进行筛选，选定某一个 Consequent，此处选择“回归分析、应用随机过程”得到多条规则如表 13 所示。

Table 13. Association rules based on educational data with “Regression Analysis, Applying Random Processes” as the posterior term
表 13. 基于教育数据后项为“回归分析、应用随机过程”的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
常微分方程、数学分析、数值分析、概率论与数理统计、大学物理	回归分析、应用随机过程	0.0639	0.6667	7.3000	2.7260	0.6522
数值分析、常微分方程、数学分析、概率论与数理统计	回归分析、应用随机过程	0.0639	0.6364	6.9682	2.4989	0.6250
数值分析、概率论与数理统计、大学物理、数学分析	回归分析、应用随机过程	0.0639	0.6364	6.9682	2.4989	0.6250
	…					
大学物理、数学分析	回归分析、应用随机过程	0.0731	0.5161	5.6516	1.8779	0.5152
大学物理、常微分方程、数学分析	回归分析、应用随机过程	0.0639	0.5000	5.4750	1.8174	0.5000
数值分析、常微分方程、大学物理	回归分析、应用随机过程	0.0639	0.5000	5.4750	1.8174	0.5000

将表 13 按照提升度、前项集降序排列，选取前五条规则结果如表 14 所示。将表 13 按照确信度、前项集降序排列，选取前五条规则结果如表 15 所示。

从表 6 进行筛选，随机选定某一个 Confidence，此处选择 Confidence = 0.875 得到多条规则如表 16 所示。

Table 14. Arrange the top 5 association rules of Table 13 in descending order of lift and antecedents
表 14. 将表 13 按照提升度、前项集降序排列的前 5 条关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
常微分方程、数学分析、数值分析、概率论与数理统计、大学物理	回归分析、应用随机过程	0.0639	0.6667	7.3000	2.7260	0.6522
数值分析、概率论与数理统计、大学物理、数学分析	回归分析、应用随机过程	0.0639	0.6364	6.9682	2.4989	0.6250
数值分析、常微分方程、数学分析、	回归分析、应用随机过程	0.0639	0.6364	6.9682	2.4989	0.6250
数值分析、常微分方程、大学物理、概率论与数理统计	回归分析、应用随机过程	0.0639	0.6087	6.6652	2.3222	0.6000
数值分析、概率论与数理统计、数学分析	回归分析、应用随机过程	0.0639	0.5833	6.3875	2.1808	0.5769

Table 15. The top 5 association rules Arranged from Table 13 in descending order of conviction and antecedents
表 15. 将表 13 按照确信度、前项集降序排列的前 5 条关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
常微分方程、数学分析、数值分析、概率论与数理统计、大学物理	回归分析、应用随机过程	0.0639	0.6667	7.3000	2.7260	0.6522
数值分析、概率论与数理统计、大学物理、数学分析	回归分析、应用随机过程	0.0639	0.6364	6.9682	2.4989	0.6250
数值分析、常微分方程、数学分析、概率论与数理统计	回归分析、应用随机过程	0.0639	0.6364	6.9682	2.4989	0.6250
数值分析、常微分方程、大学物理、概率论与数理统计	回归分析、应用随机过程	0.0639	0.6087	6.6652	2.3222	0.6000
数值分析、概率论与数理统计、数学分析	回归分析、应用随机过程	0.0639	0.5833	6.3875	2.1808	0.5769

Table 16. Association rules based on educational data with the confidence level of 0.875
表 16. 基于教育数据置信度为 0.875 的关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
大学生就业指导、空间解析几何	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学生就业指导、数学分析	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学生就业指导、高等代数	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
…						
高等代数、算法与数据结构、空间解析几何、概率论与数理统计、大学物理	数学实验	0.0639	0.8750	4.6738	6.5023	0.8333
高等代数、空间解析几何、数学分析、数值分析、概率论与数理统计、复变函数论	数学实验	0.0639	0.8750	4.6738	6.5023	0.8333
高等代数、空间解析几何、数学分析、概率论与数理统计、大学物理、复变函数论	数学实验	0.0639	0.8750	4.6738	6.5023	0.8333

将表 16 按照提升度、前项集降序排列，选取前五条条规则如表 17 所示。将表 16 按照确信度、前项集降序排列，选取前五条条规则如表 18 所示。

Table 17. The top 5 association rules arranged from Table 16 in descending order of lift and antecedents
表 17. 将表 16 按照提升度、前项集降序排列的前 5 条关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
大学物理、大学生就业指导、数学分析	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学物理、大学生就业指导、常微分方程、数学分析	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学生就业指导、数学分析	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学生就业指导、空间解析几何	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学生就业指导、高等代数	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333

Table 18. The top 5 association rules arranged from Table 16 in descending order of conviction and antecedents
表 18. 将表 16 按照确信度、前项集降序排列的前 5 条关联规则

前项	后项	支持度	置信度	提升度	确信度	Laplace
大学物理、大学生就业指导、数学分析	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学物理、大学生就业指导、常微分方程、数学分析	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学生就业指导、数学分析	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学生就业指导、空间解析几何	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333
大学生就业指导、高等代数	抽样调查	0.0639	0.8750	6.1815	6.8676	0.8333

综合表 14、表 15 和表 17、表 18 这两组表格进行观察，发现通过两种方法所选择到的规则是相同的，因此结论得到印证。也就是在假设情况的前提条件下，利用确信度或者提升度所筛选到的规则是一致的。

6. 结论

本文关于关联规则分析中两个重要的兴趣度量 Lift 和 Conviction 进行了研究，找到了二者的函数关系和单调性变化情况等。最后还从教育数据的关联规则挖掘的过程中，对得到的函数关系定理和结论进行了验证，且实验验证结果与所提出的定理一致。因此，可以认为用 Lift $(X \to Y)$或者 Conviction $(X \to Y)$找到的关联规则是一模一样的。未来，针对这二种兴趣度量，研究者们只需要选择其中之一进行使用就可以了。此外，如果要研究 X 和 Y 的双向关系，选择 Conviction 的话，可以从 Conviction $(X \to Y)$和 Conviction $(Y \to X)$两个方面研究，和 Lift 是一样的。

参考文献

[1] Rakesh, A. and Ramakrishnan, S. (1994) Fast Algorithms for Mining Association Rules in Large Databases. Morgan Kaufmann Publishers Inc., 487-499.

[2] Brin, S., Motwani, R., Ullman, J.D. and Tsur, S. (1997) Dynamic Itemset Counting and Implication Rules for Market Basket Data. *ACM SIGMOD Record*, **26**, 255-264. https://doi.org/10.1145/253262.253325

[3] Wong, M., Sze-To, H., Lo, L., Chan, T. and Leung, K. (2015) Discovering Binding Cores in Protein-DNA Binding Using Association Rule Mining with Statistical Measures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **12**, 142-154. https://doi.org/10.1109/tcbb.2014.2343952

[4] Xiang, Y., Shuai, C., Li, Y. and Zhang, Y. (2022) Information Reconstruction of Student Management Work Based on Association Rules Mining. *Computational Intelligence and Neuroscience*, **2022**, Article ID: 2318515. https://doi.org/10.1155/2022/2318515

[5] Shrestha, A., Zikos, D. and Fegaras, L. (2021) An Annotated Association Mining Approach for Extracting and Visualizing Interesting Clinical Events. *International Journal of Medical Informatics*, **148**, Article ID: 104366. https://doi.org/10.1016/j.ijmedinf.2020.104366

[6] Hu, K., Qiu, L., Zhang, S., Wang, Z. and Fang, N. (2023) An Animal Dynamic Migration Optimization Method for Directional Association Rule Mining. *Expert Systems with Applications*, **211**, Article ID: 118617. https://doi.org/10.1016/j.eswa.2022.118617

[7] Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H. and Yamaguchi, T. (2004) Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis. 8*th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Pisa, 20-24 September 2004, 362-373. https://doi.org/10.1007/978-3-540-30116-5_34

[8] Brin, S., Motwani, R. and Silverstein, C. (1997) Beyond Market Baskets: Generalizing Association Rules to Correlations. *ACM SIGMOD Record*, **26**, 265-276. https://doi.org/10.1145/253260.253327

[9] Azevedo, P.J. and Jorge, A.M. (2007) Comparing Rule Measures for Predictive Association Rules. In: *Machine Learning*: *ECML* 2007, Springer, 510-517. https://doi.org/10.1007/978-3-540-74958-5_47

[10] Lenca, P., Meyer, P., Vaillant, B. and Lallich, S. (2008) On Selecting Interestingness Measures for Association Rules: User Oriented Description and Multiple Criteria Decision Aid. *European Journal of Operational Research*, **184**, 610-626. https://doi.org/10.1016/j.ejor.2006.10.059

[11] 姜盛彬. 数据挖掘中基于兴趣度的关联规则研究[D]: [硕士学位论文]. 长沙: 湖南师范大学, 2016.

[12] Lobo, D. (2014) Association Rules: Normalizing the Lift. 9*th International Conference on Digital Information Management* (*ICDIM* 2014), Phitsanulok, 29 September-1 October 2014, 151-155. https://doi.org/10.1109/icdim.2014.6991393

[13] Ordonez, C. and Zhao, K. (2011) Evaluating Association Rules and Decision Trees to Predict Multiple Target Attributes. *Intelligent Data Analysis*, **15**, 173-192. https://doi.org/10.3233/ida-2010-0462

[14] 邱均平, 崔腾腾, 陈仕吉. 基于聚类和关联规则的 Altmetric TOP 榜文献特征分析[J]. 现代情报, 2021, 41(9): 12-21+63.

[15] 李鑫, 史天运, 常宝, 等. 基于优化的 MsEclat 算法的铁路机车事故故障关联规则挖掘[J]. 中国铁道科学, 2021, 42(4): 155-165.

[16] 王枭翔. 基于相关兴趣度的关联规则挖掘[D]: [硕士学位论文]. 兰州: 兰州交通大学, 2014.

[17] Bao, F., Mao, L., Zhu, Y., Xiao, C. and Xu, C. (2021) An Improved Evaluation Methodology for Mining Association Rules. *Axioms*, **11**, Article No. 17. https://doi.org/10.3390/axioms11010017

[18] McNicholas, P.D., Murphy, T.B. and O'Regan, M. (2008) Standardising the Lift of an Association Rule. *Computational Statistics & Data Analysis*, **52**, 4712-4721. https://doi.org/10.1016/j.csda.2008.03.013

[19] Liu, B., Hsu, W. and Ma, Y. (1998) Integrating Classification and Association Rule Mining. *Proceedings of KDD*, Vol. 1711, 80-86.

[20] 朱晓燕, 宋擒豹. 基于排序的关联分类算法[J]. 计算机科学, 2009, 36(7): 204-207.

[21] 宋威, 高磊, 李晋宏. 一种基于闭项集的无冗余关联规则挖掘方法[J]. 北京交通大学学报, 2009, 33(6): 91-96.

[22] Zaki, M.J., Meira Jr., W. 数据挖掘与分析概念与算法[M]. 吴诚堃, 译. 北京: 人民邮电出版社, 2017: 186-189.

[23] 李珺, 刘鹤, 朱良宽. 基于改进的 K-means 算法的关联规则数据挖掘研究[J]. 小型微型计算机系统, 2021, 42(1): 15-19.

[24] 李珺, 刘鹤, 朱良宽. 基于 Apriori 关联规则算法的草莓叶片含水状况研究[J]. 北方园艺, 2020(19): 146-151.

[25] Hussein, N., Alashqur, A. and Sowan, B. (2015) Using the Interestingness Measure Lift to Generate Association Rules. *Journal of Advanced Computer Science & Technology*, **4**, 156-162. https://doi.org/10.14419/jacst.v4i1.4398