Hans Journal of Data Mining 数据挖掘, 2024, 14(3), 162-171
Published Online July 2024 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2024.143015

Hans汉斯

关联规则挖掘中几个兴趣度量的值域研究

万　鑫*，李建军*，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8日；录用日期：2024年7月8日；发布日期：2024年7月17日

摘　要

本文旨在研究关联规则挖掘中的各种兴趣度量的值域问题。首先，详细介绍了关联规则挖掘过程中涉及的定义和支持度、置信度、确信度、提升度和Laplace测度这五种兴趣度量的定义，并通过具体例子对这些度量进行了说明和解释。然后，深入探讨了这五种兴趣度量的值域，并给出了其在数据库大小有限和接近无穷两种情况下的值域情况。此外，本文还对这些兴趣度量值域的区间端点的取值进行了细致讨论，指出了与其他研究结果的区别及其原因，并给出了严谨的数学证明和对比分析，为关联规则挖掘提供了更全面和准确的度量工具。

关键词

关联规则挖掘，兴趣度量，值域

Research on the Value Domains of Several Interest Measures in Association Rule Mining

Xin Wan*, Jianjun Li*, Yumei Li

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

Received: Jun. 8th, 2024; accepted: Jul. 8th, 2024; published: Jul. 17th, 2024

Abstract

This article aims to study the value domains problem of several interest metrics in association rule mining. Firstly, a detailed introduction was given to the definitions of five interest measures involved in the process of association rule mining, including support, confidence, conviction, lift, and Laplace measures. These measures were explained and illustrated through specific examples. Then, the value domains of these five interest measures were explored in depth, and their value

*共第一作者。

文章引用: 万鑫, 李建军, 李裕梅. 关联规则挖掘中几个兴趣度量的值域研究[J]. 数据挖掘, 2024, 14(3): 162-171.
DOI: 10.12677/hjdm.2024.143015

situations were given in two scenarios: limited database size and near infinite database size. In addition, this article also provides a detailed discussion on the values at the interval endpoints of these interest measures, pointing out the differences and reasons from other research results, and providing a more comprehensive and accurate measurement tool for association rule mining through rigorous mathematical proof and comparative analysis.

Keywords

Association Rule Mining, Interest Measures, Value Domain

1. 引言

兴趣度量是关联规则挖掘过程中必不可少的一部分，不论是传统的基于支持度构建的兴趣度量，还是在模糊关联规则和高效用项集挖掘过程中基于推广的支持度的兴趣度量，它们都为删除冗余关联规则和挖掘感兴趣的关联规则发挥了极大的作用[1] [2]。这些兴趣度量的提出为科研人员挖掘自己感兴趣的关联规则提供了便利。

在众多的兴趣度量中，支持度、置信度、提升度、确信度、Laplace 测度是公认的挖掘强关联规则最常用的兴趣度量，这些兴趣度量也被应用到工业、金融等领域的各种各样问题中[3]-[8]。然而对于兴趣度量的值域的研究还比较少，涉及的文章也是只给出值域，并没有进行相应的证明[9]-[11]。此外这些文章给出的兴趣度量值域不太一致。为了研究清楚这些兴趣度量的值域，本文严格证明了支持度、置信度、确信度、提升度、Laplace 兴趣度量的取值范围。

2. 相关兴趣度量的定义

本节将本文涉及到的定义进行数学描述，方便关联规则挖掘算法以及本文后续所提定理的证明中使用。共涉及 13 个定义，包括：项集、事务标识符集、事务、和数据库表示这四种基础名词定义；两种映射函数的描述；五种兴趣度量计算方法与数学表达式。同时将根据一个较为简单的数据集 D 进行举例便于更直观的理解这些定义，假设数据集 D 包括 6 次的交易数据，每种交易物品用一个英文字母表示，每次交易的内容分别表示为：“交易 1：$\{a, b, d, e\}$；交易 2：$\{b, c, e\}$；交易 3：$\{a, b, d, e\}$；交易 4：$\{a, b, c, e\}$；交易 5：$\{a, b, c, d, e\}$；交易 6：$\{b, c, d\}$”[12]。

定义 1 (项集) [12]若 $\mathcal{I}=\{x_1, x_2, \cdots, x_m\}$ 为一个集合，它由一组称作项的元素所构成，则集合 $X \subseteq \mathcal{I}$ 称为项集。

根据上述给出的数据集，可以看出共有 5 种交易物品，分别是：“a, b, c, d, e”。那么集合 $\mathcal{I}=\{a,b,c,d,e\}$，它的任意一个子集都可以称为一个项集，比如 $\{b\}$ 是一个一项集，$\{a,b\}$ 是一个二项集，$\{a,b,c\}$ 是一个三项集。

定义 2 (事务标识符集) [12]若 $\mathcal{T}=\{t_1, t_2, \cdots, t_n\}$ 为一个由事务标识符构成的集合，则集合 $T \subseteq \mathcal{T}$ 称为一个事务标识符集。

事务标识符集是由一系列的事务标识符构成，以数据集 D 为例，一个事务标识符可以是一个购物清单的编号或者是人为给定的一系列不重复序号。那么，事务标识符集就是一个购物清单的编号或者是人

为给定的一系列不重复序号的集合。假设将购物清单的编号作为数据集 D 中的事务标识符，则有 $t_1=1, t_2=2, t_3=3, \cdots, t_6=6$，即事务标识符集 $\mathcal{T}=\{1,2,3,4,5,6\}$。

定义 3 (事务) [12] 一个形如 $\langle t, X\rangle$ 的元组称为事务，其中 $t\in\mathcal{T}$ 是一个独一的标识符，X 是一个项集。

以数据库 D 为例，每一条交易可认为是一条事务，即 $\langle 1,\{a,b,d,e\}\rangle$，$\langle 2,\{b,c,e\}\rangle$，$\langle 3,\{a,b,d,e\}\rangle$，$\langle 4,\{a,b,c,e\}\rangle$，$\langle 5,\{a,b,c,d,e\}\rangle$，$\langle 6,\{b,c,d\}\rangle$，这 6 条交易，总共 6 个事务。

定义 4 (数据库表示) [12]一个二元数据库 D 表示了事务标识符和项集之间的二元关系，即 $D\subseteq\mathcal{T}\times\mathcal{I}$。

利用定义 1~3 的名词解释可以将一个数据集进行数据库表示。将数据集 D 进行事务数据库表示的过程，就是将定义 3 中所给出的所有事务进行二维表格展示，如表 1 所示。

Table 1. Transaction database representation of dataset D
表 1. 数据集 D 的事务数据库表示

事务标识符	交易物品
1	$\{a,b,d,e\}$
2	$\{b,c,e\}$
3	$\{a,b,d,e\}$
4	$\{a,b,c,e\}$
5	$\{a,b,c,d,e\}$
6	$\{b,c,d\}$

定义 5 (项集函数) [12]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：

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

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。

项集函数 $i(T)$ 是事务标识符集 T 中每个事务标识符 t_i 包含的公共项的集合。例如：$i(1)=\{a,b,d,e\}$，$i(2,3)=\{b,e\}$，$i(4,5,6)=\{b,c\}$，$i(1,2,3,6)=\{b\}$。这里的函数自变量没有采用集合的形式书写主要是为了书写方便以及形式上的美观，实际应该型如：$i(\{1\})$，$i(\{2,3\})$，$i(\{4,5,6\})$，$i(\{1,2,3,6\})$……下方标识符集函数的书写同样遵循这种规则。

定义 6 (标识符集函数) [12]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：

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

标识符集函数 $t(X)$ 是由一系列事务标识符所构成的集合，这些事务标识符需要满足以下条件，即其对应的项集应包含项集 X 中所有的项。例如：$t(a)=\{1,3,4,5\}$，$t(a,b)=\{1,3,4,5\}$，$t(a,b,c)=\{4,5\}$……

定义 7 (支持度) [12]一个项集 X 的支持度为：

$$Support(X)=\frac{\left|\{t\,|\,\langle t,i(t)\rangle\in D\wedge X\subseteq i(t)\}\right|}{|D|}=\frac{|t(X)|}{|D|}=P(X), \tag{1}$$

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

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 $P(abc)$。从表 1 中能够很容易的求出 $P(abc)=1/3$，则 $Support(X)=1/3$。

定义 8 (置信度) [12] X 发生的前提下，Y 发生的概率称为置信度：

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

定义 9 (提升度) [7] 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：

$$Lift(X\to Y)=\frac{Support(X\to Y)}{Support(X)Support(Y)}=\frac{Confidence(X\to Y)}{Support(Y)}=\frac{P(XY)}{P(X)P(Y)}. \tag{3}$$

定义 10 (确信度) [13]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：

$$Conviction(X\to Y)=\frac{1-Support(Y)}{1-Confidence(X\to Y)}=\frac{P(\bar{Y})}{P(\bar{Y}\mid X)}=\frac{1-P(Y)}{1-\frac{P(XY)}{P(X)}}. \tag{4}$$

定义 11 (拉普拉斯测度) [10]拉普拉斯测度是一个考虑了支持度的置信度估计，定义为：

$$Laplace(X\to Y)=\frac{|t(XY)|+1}{|t(X)|+2}=\frac{Support(X\to Y)+1/N}{Support(X)+2/N}=\frac{P(XY)+1/N}{P(X)+2/N}. \tag{5}$$

置信度、提升度、确信度以及拉普拉斯测度都是在支持度的基础上，利用前项、后项、前项后项共现以及它们对立事件的支持度进行计算的。

以关联规则 $\{a,b\}\to\{c,e\}$ 为例计算上述四个兴趣度量，首先需要计算项集 $\{a,b\}$，$\{c,e\}$ 和关联规则 $\{a,b\}\to\{c,e\}$ 的支持度，经计算 $Support(\{a,b\})=2/3$，$Support(\{c,e\})=1/3$，$Support(\{a,b\}\to\{c,e\})=1/3$。进而再计算这四个兴趣度量，根据公式(2)~公式(5)计算得到：

$$Confidence(\{a,b\}\to\{c,e\})=\frac{Support(\{a,b\}\to\{c,e\})}{Support(\{a,b\})}=\frac{1/3}{2/3}=\frac{1}{2},$$

$$Lift(X\to Y)=\frac{Confidence(\{a,b\}\to\{c,e\})}{Support(\{c,e\})}=\frac{1/2}{1/3}=\frac{3}{2},$$

$$Conviction(X\to Y)=\frac{1-Support(\{c,e\})}{1-Confidence(\{a,b\}\to\{c,e\})}=\frac{1-1/3}{1-1/2}=\frac{4}{3},$$

$$Laplace(X\to Y)=\frac{Support(\{a,b\}\to\{c,e\})+1/6}{Support(\{a,b\})+2/6}=\frac{1/3+1/6}{2/3+2/6}=\frac{1}{2}.$$

3. 相关兴趣度量的值域

本节将给出五种兴趣度量值域，同时还罗列出不同文章所给出的值域并在表 1 中进行对比分析。

定理 1 (支持度的值域)设数据库的大小$|D|$等于 N，项集 X 在数据库 D 中出现的次数为 N_X，项集 XY 在数据库 D 中出现的次数为 N_{XY}。那么支持度的值域为：$Support(X)\in[0,1]$，$Support(X\to Y)\in[0,1]$。

证明：

因为 $N_X\in[0,N]$，所以根据支持度定义(1)有 $Support(X)=\frac{N_X}{N}\in[0,1]$。

$Support(X\to Y)\in[0,1]$同理。

定理 2 (置信度的值域)设数据库的大小$|D|$等于 N。那么置信度的值域为：$Confidence(X\to Y)\in[0,1]$。

证明：

由置信度定义(2)可知 $Confidence(X\to Y)=\frac{P(XY)}{P(X)}=\frac{\frac{N_{XY}}{N}}{\frac{N_X}{N}}=\frac{N_{XY}}{N_X}$。又因为 $0\le N_{XY}\le N_X$，所以 $Confidence(X\to Y)\in[0,1]$。

定理 3 (确信度的值域)设数据库的大小$|D|$等于 N，且 $P(X)\ne 0$，$P(X)\ne 1$，$P(Y)\ne 0$，$P(Y)\ne 1$，那么确信度的值域为：$Conviction(X\to Y)\in\left[\frac{1}{N},\infty\right)$，当 $N\to\infty$ 时 $Conviction(X\to Y)\in(0,\infty)$。

证明：

为了书写清晰，不妨设 $P(X)=\alpha$，$P(Y)=\beta$，$P(XY)=\gamma$，其中 α，β 为常数，γ 为变量。

因为 $P(XY)$ 是项集 X 与 Y 的联合概率，所以可以确定 γ 的取值范围为：

$$0\le\gamma\le\min\{\alpha,\beta\}. \tag{6}$$

首先根据确信度的定义(4)可知计算式为：

$$Conviction(X\to Y)=\frac{1-\beta}{1-\frac{\gamma}{\alpha}}. \tag{7}$$

然后确定 $Conviction(X\to Y)$ 的连续性与单调性。

1) 连续性

$Conviction(X\to Y)$ 存在一个间断点为 $\gamma=\alpha$ 处。

2) 单调性

为了确定单调性，对 $Conviction(X\to Y)$ 求一阶导数：

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}Conviction(X\to Y)=\frac{\alpha(1-\beta)}{(\alpha-\gamma)^2}.$$

因为 $0\le\alpha\le 1$，$0\le\beta\le 1$，所以在区间 $0\le\gamma\le\alpha$ 和 $\alpha\le\gamma\le\min\{\alpha,\beta\}$ 上 $\frac{\mathrm{d}}{\mathrm{d}\gamma}Conviction(X\to Y)\ge 0$，因此 $Conviction(X\to Y)$ 在间断点两侧分别是关于 γ 的单调递增函数。

根据公式(6)可以知道 $0\le\gamma\le\min\{\alpha,\beta\}\le\alpha$，结合公式(7)，那么 $Conviction(X\to Y)$ 的取值范围就是：

$$1-\beta=\frac{1-\beta}{1-\frac{0}{\alpha}}\le Conviction\left(X\to Y\right)\le\frac{1-\beta}{1-\frac{\min\left\{\alpha,\beta\right\}}{\alpha}}. \tag{8}$$

3) 考虑$1-\beta$的最小值

因为$\max\beta=\frac{N-1}{N}$可以得到：

$$\min\left\{1-\beta\right\}=\frac{1}{N}. \tag{9}$$

4) 考虑$\frac{1-\beta}{1-\frac{\min\left\{\alpha,\beta\right\}}{\alpha}}$的最大值

① 当$\alpha\le\beta$时，$\frac{1-\beta}{1-\frac{\min\left\{\alpha,\beta\right\}}{\alpha}}=\frac{1-\beta}{1-\frac{\alpha}{\alpha}}=\infty$。

② 当$\alpha>\beta$时，$\frac{1-\beta}{1-\frac{\min\left\{\alpha,\beta\right\}}{\alpha}}=\frac{1-\beta}{1-\frac{\beta}{\alpha}}$。显然当$\alpha$与$\beta$非常接近时此式趋于无穷。

因此可以得到：

$$\frac{1-\beta}{1-\frac{\min\left\{\alpha,\beta\right\}}{\alpha}}\to\infty. \tag{10}$$

将公式(9)，(10)代入到公式(8)中可以得到$Conviction\left(X\to Y\right)$的值域为：

$$Conviction\left(X\to Y\right)\in\left[\frac{1}{N},\infty\right).$$

注(假设条件的解释)：当$P\left(X\right)=0$或$P\left(X\right)=1$时，并不能根据数据判断出X对Y的影响。当$P\left(Y\right)=0$或$P\left(Y\right)=1$时，同理。因此提出$P\left(X\right)\ne0$，$P\left(X\right)\ne1$，$P\left(Y\right)\ne0$，$P\left(Y\right)\ne1$的假设。

定理 4 (提升度的值域)设数据库的大小$\left|D\right|$等于N，且$P\left(X\right)\ne0$，$P\left(X\right)\ne1$，$P\left(Y\right)\ne0$，$P\left(Y\right)\ne1$，$P\left(X\right)$，那么提升度的值域为：$Lift\left(X\to Y\right)\in\left[0,N\right]$，当$N\to\infty$，$Lift\left(X\to Y\right)\in\left[0,\infty\right)$。

证明：

为了书写清晰，不妨设$P\left(X\right)=\alpha$，$P\left(Y\right)=\beta$，$P\left(XY\right)=\gamma$，其中α,β为常数γ为变量。因为$P\left(XY\right)$是项集X与Y的联合概率，所以可以确定γ的取值范围为：

$$0\le\gamma\le\min\left\{\alpha,\beta\right\}. \tag{11}$$

根据提升度的定义(3)可知计算式为：

$$Lift\left(X\to Y\right)=\frac{P\left(XY\right)}{P\left(X\right)P\left(Y\right)}=\frac{\gamma}{\alpha\beta}. \tag{12}$$

$Lift\left(X\to Y\right)$显然是关于γ的连续单调递增函数。根据公式(11)和公式(12)可知$Lift\left(X\to Y\right)$的取值范围就是：

$$0=\frac{0}{\alpha\beta}\le Lift\left(X\to Y\right)\le\frac{\min\left\{\alpha,\beta\right\}}{\alpha\beta}=\min\left\{\frac{1}{\alpha},\frac{1}{\beta}\right\}. \tag{13}$$

因为 $\min\alpha=\frac{1}{N}$， $\min\beta=\frac{1}{N}$，所以

$$\min\left\{\frac{1}{\alpha},\frac{1}{\beta}\right\}=N. \tag{14}$$

将公式(14)代入公式(13)中得到提升度的值域为：

$$Lift(X\to Y)\in[0,N].$$

定理 5 (Laplace 测度的值域)设数据库的大小 $|D|$ 等于 N，且 $P(X)\neq 0$，$P(X)\neq 1$，$P(Y)\neq 0$，$P(Y)\neq 1$。那么 Laplace 测度的值域为：$Laplace(X\to Y)\in\left[\frac{1}{N+1},\frac{N}{N+1}\right]$，当 $N\to\infty$，$Laplace(X\to Y)\in(0,1)$。

证明：

为了书写清晰，不妨设 $P(X)=\alpha$，$P(Y)=\beta$，$P(XY)=\gamma$，其中 α,β 为常数 γ 为变量。因为 $P(XY)$ 是项集 X 与 Y 的联合概率，所以可以确定 γ 的取值范围为：

$$0\le\gamma\le\min\{\alpha,\beta\}. \tag{15}$$

根据 $Laplace(X\to Y)$ 的定义(5)可知计算式为：

$$Laplace(X\to Y)=\frac{P(XY)+\frac{1}{N}}{P(X)+\frac{2}{N}}=\frac{\gamma+\frac{1}{N}}{\alpha+\frac{2}{N}}. \tag{16}$$

$Laplace(X\to Y)$ 显然是关于 γ 的连续单调递增函数。根据公式(15)以及公式(16)，可知 $Laplace(X\to Y)$ 的取值范围就是：

$$\frac{\frac{1}{N}}{\alpha+\frac{2}{N}}\le Laplace(X\to Y)\le\frac{\min\{\alpha,\beta\}+\frac{1}{N}}{\alpha+\frac{2}{N}}. \tag{17}$$

1) 考虑 $\frac{\frac{1}{N}}{\alpha+\frac{2}{N}}$ 的最小值：

因为 $\max\alpha=\frac{N-1}{N}$，所以可以得到：

$$\min\left\{\frac{\frac{1}{N}}{\alpha+\frac{2}{N}}\right\}=\frac{1}{N+1}. \tag{18}$$

2) 考虑 $\frac{\min\{\alpha,\beta\}+\frac{1}{N}}{\alpha+\frac{2}{N}}$ 的最大值：

① 当 $\alpha\le\beta$ 时，$\frac{\min\{\alpha,\beta\}+\frac{1}{N}}{\alpha+\frac{2}{N}}=\frac{\alpha+\frac{1}{N}}{\alpha+\frac{2}{N}}=1-\frac{\frac{1}{N}}{\alpha+\frac{2}{N}}$，要使 $\frac{\min\{\alpha,\beta\}+\frac{1}{N}}{\alpha+\frac{2}{N}}$ 达到最大，即使 α 取得最

大，即 $\max\alpha=\frac{N-1}{N}$，此时 $\frac{\min\{\alpha,\beta\}+\frac{1}{N}}{\alpha+\frac{2}{N}}=\frac{N}{N+1}$。

② 当 $\alpha>\beta$ 时，$\frac{\min\{\alpha,\beta\}+\frac{1}{N}}{\alpha+\frac{2}{N}}=\frac{\beta+\frac{1}{N}}{\alpha+\frac{2}{N}}$，因为 $\alpha>\beta$，所以 $\frac{\beta+\frac{1}{N}}{\alpha+\frac{2}{N}}<\frac{\alpha+\frac{1}{N}}{\alpha+\frac{2}{N}}$，即 $\frac{\min\{\alpha,\beta\}+\frac{1}{N}}{\alpha+\frac{2}{N}}<\frac{N}{N+1}$。

综合①，②所述，可以得到：

$$\max\left\{\frac{\min\{\alpha,\beta\}+1}{\alpha+2}\right\}=\frac{N}{N+1}. \tag{19}$$

将公式(18)，(19)代入到公式(17)中可以得到 $Laplace(X\to Y)$ 的值域为：

$$Laplace(X\to Y)\in\left[\frac{1}{N+1},\frac{N}{N+1}\right].$$

将定理 1~定理 5 所给出的不同兴趣度量值域与另外两位作者的文章所给出的值域进行综合展示，如表 2 所示。

Table 2. Comparison of interest measures' value domain
表 2. 兴趣度量值域对比表

兴趣度量		P.J. Azevedo and A.M. Jorge [9]	P. Lenca, P. Meyer 等[10]	本文	
				N_X,N_Y 已知	N_X,N_Y 未知
支持度	$\lvert D\rvert\to\infty$	----	----	----	$[0,1]$
	$\lvert D\rvert=N$	----	$[0,N_X/N]$	----	$[0,1]$
置信度	$\lvert D\rvert\to\infty$	$[0,1]$	----	----	$[0,1]$
	$\lvert D\rvert=N$	----	$[0,1]$	----	$[0,1]$
确信度	$\lvert D\rvert\to\infty$	$\left[\frac{1}{2},+\infty\right)$	----	----	$(0,+\infty)$
	$\lvert D\rvert=N$	----	$[N_{\bar{Y}}/N,+\infty)$	$[N_{\bar{Y}}/N,+\infty)$	$[1/N,+\infty)$
提升度	$\lvert D\rvert\to\infty$	$[0,+\infty)$	----	----	$[0,+\infty)$
	$\lvert D\rvert=N$	----	$[0,N/N_Y]$	$\left[0,\max\left\{\frac{N}{N_X},\frac{N}{N_Y}\right\}\right]$	$[0,N]$
Laplace	$\lvert D\rvert\to\infty$	$[0,1)$	----	----	$(0,1)$
	$\lvert D\rvert=N$	----	$\left[\frac{1}{N_X+2},\frac{N_X+1}{N_X+2}\right]$	$\left[\frac{1}{N_X+2},\frac{\min\{N_X,N_Y\}+1}{N_X+2}\right]$	$\left[\frac{1}{N+1},\frac{N}{N+1}\right]$

注：表中 N 表示某一数据库实际包含事务个数；N_X 表示数据库 D 中包含项集 X 的事务个数，也就是项集 X 在数据库中出现的次数，N_Y 表示数据库 D 中包含项集 Y 的事务个数，即项集 Y 在数据库中出现的次数；$N_{\bar{Y}}$ 表示项集 Y 在数据库中没有出现的次数，也就是数据库实际包含事务个数减去项集 Y 在数据库中出现的次数，即 $N_{\bar{Y}}=N-N_Y$。

从表 2 中可以发现，本文给出了数据库大小是常数以及数据库大小趋近于无穷的情况下兴趣度量的值域。本文所给出的兴趣度量的值域与 P.J. Azevedo, A.M. Jorge [9]的文章和 P. Lenca, P. Meyer 等人[10]的文章存在的不同主要有两方面，在表格中使用红色和蓝色分别标出。红色部分是与 P.J. Azevedo 和 A.M. Jorge [9]的文章在值域的下界上存在差异：本文认为确信度的最小值可以小于 0.5；Laplace 的下界不能等于 0。

蓝色部分是与 P. Lenca, P. Meyer 等人[10]在 N_X 和 N_Y 已知的情况下，兴趣度量的值域上界存在差异，出现这些差异的主要原因是，该作者认为项集 X 在数据库中出现的次数比项集 Y 在数据库中出现的次数多，即 $N_X > N_Y$。因为关联规则 $X \to Y$ 所表达的是 X 的出现引起 Y 的出现，因此直观上该作者这种理解是正确的，但是在关联规则挖掘过程并没有做出这一假设，所挖掘的关联规则也并没有排除符合 $N_Y > N_X$ 这种情况的规则。本文也正式去掉了这一假设所给出的定理 1~定理 5。

4. 结论

本文从关联规则中的各种兴趣度量入手，研究了兴趣度量的值域，综合多方面考量给出了支持度、置信度、确信度、提升度与 Laplace 测度这五种兴趣度量在数据库大小是否接近无穷的两种情况下的值域以及严谨的数学证明过程，并且还与另外两篇文章所给出的兴趣度量的值域作对比，并且解释了出现差别的原因。综上所述，本文通过严谨的数学证明和综合分析，深入探讨了支持度、置信度、确信度、提升度与 Laplace 测度在不同数据库大小条件下的值域，提供了更为全面和准确的值域证明，弥补了以往研究中的不足之处。此外，不仅为关联规则研究提供了坚实的理论基础，也为实际应用中的规则评估提供了有力的支持。

参考文献

[1] 郭瑞, 钱晓东. 基于一阶谓词公式去除商务数据冗余关联规则的研究[J]. 计算机工程与科学, 2017, 39(3): 593-598.

[2] 翟悦, 秦放. 基于概念格的无冗余关联规则提取算法[J]. 计算机应用与软件, 2015, 32(4): 46-49+66.

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

[4] Ordonez, C. (2006) Comparing Association Rules and Decision Trees for Disease Prediction. *Proceedings of the International Workshop on Healthcare Information and Knowledge Management*, Association for Computing Machinery, 17-24. https://doi.org/10.1145/1183568.1183573

[5] 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.

[6] 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.

[7] 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.

[8] Bao, F.G., Mao, L.H., Zhu, Y.L., *et al.* (2022) An Improved Evaluation Methodology for Mining Association Rules. *Axioms*, **11**, 2-17. https://doi.org/10.3390/axioms11010017

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

[10] Lenca, P., Meyer, P., Vaillant, B., *et al.* (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] 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

[12] Mohammed J. Zaki, Wagner Meira Jr. 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.

[13] Brin, S., Motwani, R, Ullman, J.D., *et al.* (2001) Dynamic Itemset Counting and Implication Rules for Market Basket Data. *ACM SIGMOD Record*, **26**, 255-264. https://doi.org/10.1145/253262.253325