

# 对最小置信度门限的置疑<sup>\*</sup>

马占欣<sup>1,2</sup> 王新社<sup>1</sup> 黄维通<sup>3</sup> 陆玉昌<sup>2</sup>

(周口师范学院计算机科学系 河南周口 466001)<sup>1</sup>

(清华大学智能系统与技术国家重点实验室 北京 100084)<sup>2</sup>

(清华大学计算机科学与技术系 北京 100084)<sup>3</sup>

**摘要** 在关联规则挖掘中,通常使用最小支持度和最小置信度两个门限来衡量一条规则是不是一个强规则。本文对最小置信度这个参数的实际意义,从理论和实践上进行了分析研究和探讨,发现使用最小置信度进行限制不仅所挖掘出的规则质量较低,还有可能遗漏一些具有重要价值的规则,进一步提出提升率比置信度更能反映实际情况,在关联规则挖掘中改用最小支持度和最小提升率作为衡量准则,其结论更加准确,意义也更明确。

**关键词** 数据挖掘,关联规则,兴趣度,置信度,提升率

## Doubts about Min Confidence Threshold

MA Zhan-Xin<sup>1,2</sup> WANG Xin-She<sup>1</sup> HUANG Wei-Tong<sup>3</sup> LU Yu-Chang<sup>2</sup>

(Department of Computer Science, Zhoukou Normal University, Zhoukou, Henan 466001)<sup>1</sup>

(State Key Lab of Intelligent System and Technology, Tsinghua University, Beijing 100084)<sup>2</sup>

(Department of Computer Sciences and Techonolodge, Tsinghua University, Beijing 100084)<sup>3</sup>

**Abstract** The two thresholds: min support and min confidence are often used to evaluate if a rule is a strong rule or not in association rules mining. The present paper analyzes and explores the practical significance of min confidence both theoretically and practically. It finds that not only the quality of the mined rules is comparatively low but also some important rules are probably missed out if min confidence is used to restrict. Different from previous research, this paper proposes that upgrade rate can reflect the actuality more accurately than min confidence and the result will be more accurate and clearer if min support and min upgrade are taken as the weighing rule in association rules mining.

**Keywords** Data mining, Association rules, Interest, Confidence, Upgrade rate

## 1 传统关联规则挖掘中的两个基本参数及存在的问题

关联规则挖掘是数据挖掘中一个重要的研究方向。关联规则挖掘的目标是发现数据集中项集之间有趣的关联关系。其核心是要挖掘出这样的规则:在数据集中,项集 A 的出现会影响项集 B 的出现(诱导项集 B 的出现或抑制项集 B 的出现)。传统的关联规则挖掘中,只考虑正向的关系,即要挖掘出“项集 A 的出现会诱导项集 B 的出现”这样的关联规则。

为了挖掘出真正有意义的关联规则,通常引入两个重要的限制参数,最小支持度阈值(min-sup)和最小置信度阈值(min-conf)<sup>[1]</sup>。

支持度的定义为:

$$\text{Support}(A \Rightarrow B) = |A \cup B| / |D|$$

其中 $|A \cup B|$ 表示在数据集中项集 A 和项集 B 同时出现的事务数, $|D|$ 表示数据集中总的事务数。支持度也就是项集 A 和项集 B 同时出现的概率。

置信度的定义为:

$$\text{Confidence}(A \Rightarrow B) = \text{Support}(A \cup B) / \text{Support}(A)$$

置信度实质上是在项集 A 出现的条件下,项集 B 的出现概率。即项集 A 出现时,项集 B 出现的条件概率。

支持度描述了所挖掘出的关联规则的统计重要性,还可以理解为该条规则能够适用的广泛程度。支持度越高,说明

在整个数据集中符合该条规则的事务所占的比例越高。如果一条规则的支持度非常低,则表示符合这条规则的事务在整个数据集中所占的比例非常低,不具备统计意义,或说是没有可利用的价值。因此,在关联规则挖掘中要设置一个最小支持度阈值。对小于最小支持度的关联规则因其利用价值太低而被忽略,这就是最小支持度限制的最本质的意义。

与支持度相对应,置信度的本意是用于衡量所挖掘出的关联规则的可信程度,即“置信度越高,其可信程度也越高”。与对支持度的限制类似,在关联规则挖掘中同样引入了最小置信度阈值。对于小于最小置信度的规则,因其“可信程度太低”而被舍弃。

实际上,置信度并不能很好地描述规则的可信程度,置信度高低并不完全等于其可信赖程度的高低,两者之间没有全等的关系。置信度大于最小置信度的规则不一定就可信赖,关于这一点,在文[1]中,已举出反例进行了证明。实际上,置信度小于最小置信度的规则不一定就没有意义,在本文后面会证明这一点。

目前所流行的关联规则挖掘算法,都是采用首先以最小支持度阈值为第一门限,从数据集中找出不低于最小支持度的模式—频繁项集,然后再从频繁项集中挖掘出满足最小置信度阈值的关联规则。

由于置信度不能很好地描述关联规则的可信赖程度,因此,必然出现这样的结果:最小置信度设置过高,可能会漏掉

<sup>\*</sup> 基金项目:国家自然科学基金(60473115)。马占欣 副教授,清华大学访问学者,研究方向:数据挖掘与知识发现。陆玉昌 教授,博士生导师,研究方向:数据挖掘和机器学习。

一些有意义的规则,设置过低又必然会产生大量虚假规则。虽然最小置信度作为关联规则挖掘的重要门限之一,到目前仍然被广泛引用,而实际上置信度本身在有些情况下,可能会误导关联规则的挖掘过程。

## 2 一种新的相关度计算方法—提升率

为了弥补置信度本身存在的缺陷,一般要引入“相关度”<sup>[1]</sup>或“作用度”<sup>[2]</sup>或“提升度”<sup>[3]</sup>等概念。不同的作者叫法不同,计算方法不同,但描述的问题是一致的,都是描述两个项集之间的相关程度。以作用度为例,其定义如下:

$$\text{Lift}(A \Rightarrow B) = \text{Confidence}(A \Rightarrow B) / \text{Support}(B)$$

作用度描述的是在项集 A 出现的条件下项集 B 的出现概率(条件概率)与项集 B 在整个数据集中出现的概率之比。注意,这里的  $\text{Support}(B)$  是在整个数据集中项集 B 的出现概率,而不是 A 不出现时, B 出现的概率(条件概率)。其它所有关于相关度计算的方法中也均没有引用“A 不出现时, B 出现的条件概率”这一重要参数。

在其它相关学科的研究中,当涉及到考察某一因素对一个特征数据的影响时,一般会采用实验组与对照组的概念。例如,如果要衡量某项技术对生产的影响,一般是用采用这项技术之后的结果与采用这项技术之前的结果进行比较;医学上,为了衡量某种因素(比如某种新疗法)对某种疾病的影响,通常是将患者分为实验组和对照组,最后要把实验组的结果与对照组的结果进行比较。

在作用度的计算中,分母  $\text{Support}(B)$  是项集 B 总的出现概率。 $\text{Support}(B)$  既包含项集 A 出现对项集 B 出现概率的贡献,也包含项集 A 不出现对项集 B 出现概率的贡献。这样,其所得结果与项集 A 对项集 B 的实际贡献相比被削弱了,更严重的是,这种计算方法还会造成计算结果的不对称和不完备。这一点作者已在另文证明(已被《计算机工程》录用),作者还在该文中提出采用如下公式

$$\text{Up}(A \Rightarrow B) = \frac{P(B|A)}{P(B|\bar{A})} = \frac{\text{Confidence}(A \Rightarrow B)}{\text{Confidence}(\bar{A} \Rightarrow B)}$$

来描述项集 A 与项集 B 之间的相关程度,其结果更符合实际,且其物理意义也更加明确。如此得到的结果称为提升率(Upgrade rate),其取值范围为  $[0, \infty)$ 。公式可以按下述方式解释:把总数据集按照是否包含项集 A 划分为两部分,包含项集 A 的部分不妨称为实验组,不包含项集 A 的部分称为对照组,提升率即可理解为实验组中项集 B 出现的概率与对照组中项集 B 出现的概率之比。如果两者相等,提升率为 1,说明项集 A 的存在与否对项集 B 的出现没有影响;提升率大于 1,则说明项集 A 的出现对项集 B 的出现有诱导作用,或说项集 A 与项集 B 是正相关的,提升率越大,相关程度越高;提升率小于 1,则说明项集 A 的出现对项集 B 的出现有抑制作用,或说项集 A 与项集 B 是负相关,提升率越小,负相关程度越高。

## 3 置信度的替代指标—提升率

置信度的本意是要描述关联规则的可信程度,而实际上,由于其只使用了项集 A 出现时项集 B 出现的条件概率,未与 A 不出现时 B 出现的条件概率进行比较,因此无法准确地反映项集 A 的出现对项集 B 出现的贡献,也就无法描述规则的可信程度。不仅如此,以最小置信度作为关联规则挖掘的一个门限,还有可能遗漏一些非常有价值的规则。

既然置信度不能很好地描述关联规则的可信程度,那我们为什么不另找一个更可信的参数来描述规则的可信程度呢?实际上,提升率就是一个比较好的指标。下面我们首先

通过一个例子来观察置信度和提升率对关联规则挖掘的指导作用。

例如,在事务数据库中有 10000 条交易记录,包含 A 的 4000 条,包含 B 的 1100 条,包含 C 的 5200 条,同时包含 A 和 B 的 800 条,同时包含 A 和 C 的 1600 条。

于是有

$$\text{Support}(A \cup B) = 800 / 10000 = 8\%$$

$$\text{Support}(A \cup C) = 1600 / 10000 = 16\%$$

$$\text{Confidence}(A \Rightarrow B) = 800 / 4000 = 20\%$$

$$\text{Confidence}(B \Rightarrow A) = 800 / 1100 = 73\%$$

$$\text{Confidence}(A \Rightarrow C) = 1600 / 4000 = 40\%$$

$$\text{Confidence}(C \Rightarrow A) = 1600 / 5200 = 31\%$$

假设我们以最小支持度 5%,最小置信度 30%来挖掘关联规则,显然,  $A \Rightarrow B$  不满足最小置信度而被舍去,其它几条规则既满足最小支持度又满足最小置信度而被保留。当然,如果再进一步计算相关度会发现 A 和 C 之间存在着负相关,因此,  $A \Rightarrow C$  和  $C \Rightarrow A$  这两条规则最终也被舍去。

事实上,如果不使用置信度这一参数,而使用提升率来进行判断,将得出完全相反的结论。

$$\text{Up}(A \Rightarrow B) = \frac{P(B|A)}{P(B|\bar{A})} = \frac{20\%}{5\%} = 4$$

$$\text{Up}(B \Rightarrow A) = \frac{73\%}{36\%} = 2.02$$

$$\text{Up}(A \Rightarrow C) = \frac{40\%}{60\%} = 0.67$$

$$\text{Up}(C \Rightarrow A) = \frac{31\%}{50\%} = 0.62$$

由提升率可以看出, A 出现时 B 的出现概率是 A 不出现时 B 的出现概率的 4 倍,这表明 A 的出现与否对 B 的出现影响极大,显然是一个有趣的现象。因此,  $A \Rightarrow B$  是一条有趣的规则。 $B \Rightarrow A$  的置信度远大于最小置信度,从置信度的角度看,它是一条强规则。由提升率来看,它也是一条强规则,但  $\text{Up}(B \Rightarrow A)$  只有  $\text{Up}(A \Rightarrow B)$  的 1/2。

而 A 的出现与 C 的出现之间是负相关的关系,所以  $A \Rightarrow C$  和  $C \Rightarrow A$  均是虚假的规则。

由上例数据可以看出,用最小支持度和最小置信度模式可以挖掘出 3 条规则,排除 1 条规则。而其所挖掘出的 3 条规则中,有 1 条是真规则,2 条是伪规则;所排除的 1 条规则也是真规则。这充分说明了置信度对关联规则挖掘的指导作用是不准确的,或说置信度参数并没有描述出关联规则中最本质的属性。

如果在关联规则挖掘过程中,恰当地指定最小支持度和最小提升率,即可以挖掘出有真正意义的关联规则。

下面通过一个实例来比较最小支持度、最小置信度模式所挖掘的关联规则与最小支持度、最小提升率模式所挖掘的规则的数量和质量的不同。

## 4 最小支持度、最小置信度模式与最小支持度、最小提升率模式关联规则挖掘的比较

作者使用 SQL7 自带的模拟数据,分别用最小支持度、最小置信度和最小支持度、最小提升率两种模式,对 Sales-Fact\_1998 在产品类概念层上进行关联规则挖掘。挖掘结果如下。

### 4.1 最小支持度、最小置信度模式的关联规则挖掘

在事务数据库中,同时包含两种或两种以上商品的事务总数为 31489。以最小支持度为 2%,最小置信度 30%,共挖掘出 25 条规则,如表 1 所示。

表1 用最小支持度-最小置信度模式所挖掘出的关联规则

| A $\Rightarrow$ B     | A $\cup$ B | A    | B     | Supp | Conf  |
|-----------------------|------------|------|-------|------|-------|
| 109= $\Rightarrow$ 61 | 1518       | 4378 | 10200 | 4.82 | 34.67 |
| 8= $\Rightarrow$ 61   | 698        | 2014 | 10200 | 2.22 | 34.66 |
| 39= $\Rightarrow$ 61  | 685        | 1991 | 10200 | 2.18 | 34.40 |
| 58= $\Rightarrow$ 61  | 1502       | 4388 | 10200 | 4.77 | 34.23 |
| 1= $\Rightarrow$ 61   | 852        | 2505 | 10200 | 2.71 | 34.01 |
| 66= $\Rightarrow$ 61  | 797        | 2355 | 10200 | 2.53 | 33.84 |
| 18= $\Rightarrow$ 61  | 972        | 2899 | 10200 | 3.09 | 33.53 |
| 26= $\Rightarrow$ 61  | 687        | 2052 | 10200 | 2.18 | 33.48 |
| 99= $\Rightarrow$ 61  | 2164       | 6474 | 10200 | 6.87 | 33.43 |
| 36= $\Rightarrow$ 61  | 834        | 2505 | 10200 | 2.65 | 33.29 |
| 62= $\Rightarrow$ 61  | 953        | 2863 | 10200 | 3.03 | 33.29 |
| 60= $\Rightarrow$ 61  | 834        | 2518 | 10200 | 2.65 | 33.12 |
| 83= $\Rightarrow$ 61  | 687        | 2076 | 10200 | 2.18 | 33.09 |
| 110= $\Rightarrow$ 61 | 643        | 1944 | 10200 | 2.04 | 33.08 |
| 33= $\Rightarrow$ 61  | 811        | 2453 | 10200 | 2.58 | 33.06 |
| 49= $\Rightarrow$ 61  | 676        | 2060 | 10200 | 2.15 | 32.82 |
| 90= $\Rightarrow$ 61  | 630        | 1937 | 10200 | 2.00 | 32.52 |
| 64= $\Rightarrow$ 61  | 662        | 2070 | 10200 | 2.10 | 31.98 |
| 12= $\Rightarrow$ 61  | 758        | 2375 | 10200 | 2.41 | 31.92 |
| 80= $\Rightarrow$ 61  | 661        | 2082 | 10200 | 2.10 | 31.75 |
| 101= $\Rightarrow$ 61 | 635        | 2004 | 10200 | 2.02 | 31.69 |
| 27= $\Rightarrow$ 61  | 654        | 2087 | 10200 | 2.08 | 31.34 |
| 35= $\Rightarrow$ 61  | 633        | 2021 | 10200 | 2.01 | 31.32 |
| 45= $\Rightarrow$ 61  | 1171       | 3837 | 10200 | 3.72 | 30.52 |
| 11= $\Rightarrow$ 61  | 1307       | 4328 | 10200 | 4.15 | 30.20 |

通过 Lift 检验,去掉其中 Lift $<1$  的 8 条规则,还剩余 17 条规则。从所挖掘出的规则可以看出,所有规则均是从其它商品类导出代号为 61 的商品类,而代号 61 的商品是事务数据库中出现频率最高的。之所以没有出现其它类型的关联规则,是因为其它商品类的出现频率较低,置信度自然也较低,从而被忽略。由此也可以看出,采用最小支持度、最小置信度模式比较倾向于挖掘出后项为出现频率较高的项集的关联规则。

#### 4.2 最小支持度-最小提升率模式的关联规则挖掘结果

以最小支持度为 2%,最小提升率 1.05,共挖掘出 23 条规则,如表 2 所示。

采用最小支持度、最小提升率模式所挖掘的规则中没有虚假规则。

#### 4.3 两种模式所挖掘的关联规则比较

两种模式所挖掘的关联规则重叠的共有 6 条。采用最小支持度、最小置信度模式发现的,而被最小支持度、最小提升率模式忽略的规则有 19 条,这 19 条规则由于其提升率太低,项集之间的相关性太弱(甚至负相关)而被忽略;采用最小支持度、最小提升率模式发现而被最小支持度、最小置信度模式忽略的规则有 17 条,则是因为置信度太低而被忽略。

特别要说明的是,用最小支持度、最小提升率模式所挖掘出的规则中,提升率最高的一条规则是 109= $\Rightarrow$ 11,其置信度为:

$$\text{Confidence}(109=\Rightarrow 11)=670/4378=15.30\%$$

远低于用最小支持度、最小置信度模式所设定的最小置信度。但是,11 号商品的出现概率情况为,当 109 号商品不出现时,其出现概率为

$$P(11|\overline{109})=(4328-670)/(31489-4378)=13.49\%$$

当 109 号商品出现时,其出现概率为

$$P(11|109)=670/4378=15.30\%$$

109 号商品出现时与不出现时相比,11 号商品的出现概率提高了 0.13 倍,从提升率的角度来看,它是整个数据集中最有价值的一条规则。当然,这个提升率仍然是比较低的。作者怀疑,整个模拟数据可能是随机生成的,因此,项集与项集之间的相关性比较弱。在实际使用最小支持度、最小提升率模式进行关联规则挖掘时,应该把最小提升率设置得相对高一些,才能挖掘出真正有意义的关联规则。

表2 用最小支持度-最小提升率模式所挖掘出的关联规则

| A $\Rightarrow$ B     | A $\cup$ B | A     | B     | supp | exalt | lift |
|-----------------------|------------|-------|-------|------|-------|------|
| 109= $\Rightarrow$ 11 | 670        | 4378  | 4328  | 2.13 | 1.13  | 1.11 |
| 11= $\Rightarrow$ 109 | 670        | 4328  | 4378  | 2.13 | 1.13  | 1.11 |
| 11= $\Rightarrow$ 58  | 670        | 4328  | 4388  | 2.13 | 1.13  | 1.11 |
| 58= $\Rightarrow$ 11  | 670        | 4388  | 4328  | 2.13 | 1.13  | 1.11 |
| 61= $\Rightarrow$ 109 | 1518       | 10200 | 4378  | 4.82 | 1.11  | 1.07 |
| 61= $\Rightarrow$ 8   | 698        | 10200 | 2014  | 2.22 | 1.11  | 1.07 |
| 61= $\Rightarrow$ 39  | 685        | 10200 | 1991  | 2.18 | 1.09  | 1.06 |
| 61= $\Rightarrow$ 58  | 1502       | 10200 | 4388  | 4.77 | 1.09  | 1.06 |
| 8= $\Rightarrow$ 61   | 698        | 2014  | 10200 | 2.22 | 1.08  | 1.07 |
| 61= $\Rightarrow$ 1   | 852        | 10200 | 2505  | 2.71 | 1.08  | 1.05 |
| 99= $\Rightarrow$ 18  | 634        | 6474  | 2899  | 2.01 | 1.08  | 1.06 |
| 109= $\Rightarrow$ 61 | 1518       | 4378  | 10200 | 4.82 | 1.08  | 1.07 |
| 18= $\Rightarrow$ 99  | 634        | 2899  | 6474  | 2.01 | 1.07  | 1.06 |
| 39= $\Rightarrow$ 61  | 685        | 1991  | 10200 | 2.18 | 1.07  | 1.06 |
| 58= $\Rightarrow$ 61  | 1502       | 4388  | 10200 | 4.77 | 1.07  | 1.06 |
| 61= $\Rightarrow$ 66  | 797        | 10200 | 2355  | 2.53 | 1.07  | 1.04 |
| 99= $\Rightarrow$ 44  | 645        | 6474  | 2973  | 2.05 | 1.07  | 1.06 |
| 44= $\Rightarrow$ 99  | 645        | 2973  | 6474  | 2.05 | 1.06  | 1.06 |
| 1= $\Rightarrow$ 61   | 852        | 2505  | 10200 | 2.71 | 1.05  | 1.05 |
| 61= $\Rightarrow$ 99  | 2164       | 10200 | 6474  | 6.87 | 1.05  | 1.03 |
| 61= $\Rightarrow$ 18  | 972        | 10200 | 2899  | 3.09 | 1.05  | 1.04 |
| 61= $\Rightarrow$ 26  | 687        | 10200 | 2052  | 2.18 | 1.05  | 1.03 |
| 66= $\Rightarrow$ 61  | 797        | 2355  | 10200 | 2.53 | 1.05  | 1.04 |

**结论** 关联规则是要描述项集之间的关联关系的。如果没有比较强的相关关系,无论其置信度有多高,最终所得到的规则都有可能是虚假的规则,甚至是误导的规则。通过本文的理论分析和实验结果比较可以看出,使用最小置信度进行限制不仅所挖掘出的规则质量较低,还有可能遗漏一些具有重要价值的规则。如果在关联规则挖掘中,在设置最小支持度的基础上,再设置一个最小提升率阈值,将会使所挖掘出的规则个数有所减少,规则的质量也会有大的提高,实用性更强。

#### 参考文献

- 1 HAN Jiawei, Kamber M. Data Mining Concepts and Techniques [M]. San Francisco: Morgan Kaufmann Publishers, 2001
- 2 Dunham M H 著. 郭崇慧, 田凤占, 靳晓明译. 陆玉昌审校. 数据挖掘教程[M]. 清华大学出版社, 2005, 5: 162
- 3 Hastie T, Tibshirani R, Friedman J 著. 范明, 柴玉梅, 咎红英译. 统计学习基础—数据挖掘、推理与预测[M]. 电子工业出版社, 2004, 309
- 4 Ahmed K M, El-Makky N M, Taha Y. A note on "Beyond market basket: Generalizing association rules to correlations." [C]. SIGKDD Explorations, 2000(1): 46~48
- 5 Brin S, Motwani R, Silverstein C. Beyond market basket: Generalizing association rules to correlations[C]. In: Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data, Tucson, AZ, 1997. 265~276
- 6 罗可, 郝东妹. 挖掘有效的关联规则[J]. 小型微型计算机系统, 2005, 26(8): 1374~1379