基于客观兴趣度的关联规则评价方法①

亓文娟[1], 晏 杰[2]

[1](武夷学院 数学与计算机系, 武夷山 354300)
[2](武夷学院 团委, 武夷山 354300)

摘 要: 目前衡量和生成关联规则的主要准则是考虑支持度和置信度阈值,而在实际应用中仅按此准则来挖掘是不够的, 这主要是因为关联规则的评价标准不合理产生的. 针对关联规则评价指标进行了深入的研究, 分析了“支持度-置信度”架构的局限性, 提出了基于相关性的兴趣度的评价指标 PS 公式, 根据其数学特性指出了它的优点与不足, 为关联规则评价体系的改进奠定了理论基础.

关键词: 关联规则; 兴趣度; 支持度; 置信度; PS 公式

Objective Evaluation Method of Association Rule Interestingness

QI Wen-Juan[1], YAN Jie[2]

[1](Mathematics and Computer Science Department, Wuyi University, Wuyishan 354300, China)
[2](Youth League Committee, Wuyi University, Wuyishan 354300, China)

Abstract: Current main guidelines is to measure and generate Association rules take into account support and confidence threshold, and only in the practical application of this guideline to mining is insufficient, this is mainly because the associated rule evaluation criterion is not reasonable. This article for the associated rule evaluation conducted an in-depth study, analyzed the "support-confidence" schema limitations, presenting an interest based on correlation degree of evaluation indicators PS formula, based on its mathematical properties that has its advantages and disadvantages, laid the theoretical foundation for improvement of evaluation system of Mining Association rules.

Key words: association rules; interest measure; support; confidence; PS Formula

关联规则是数据挖掘领域成果颇丰而且比较活跃的研究分支, 最早是由 R.Agrawal 等人在 1993 年提出来的, 是对一个事物和其它事物的相互依存和相互关联的一种描述. 在包含海量数据的商业数据集上进行关联分析时, 往往会产生成千上万的关联规则, 什么是有意义的关联规则? 因此对关联规则的评价显得尤为重要, 它直接影响关联规则挖掘算法输出规则的数量和质量.

1 关联规则概述

设 I=[i1,i2,,i3,⋯ in]项的集合, 数据集 D 是事务的集合, 其中每个事务 T 是项的集合, 使得 T⊆I. 每一个事务有一个表示符, 称作 TID. 事务 T 包含一个项目集 A 当且仅当 A⊆T, 一个关联规则就是形如 A→B 的逻辑蕴涵式[1], 其中 A⊂I, B⊂I, 并且 A∩B=∅. 它表示如果项集 A 在某一事务中出现, 则必然会导致项集 B 也会在同一事务中出现. A 称为规则的先决条件(或前件), B 为规则的结果(或后件).

判断一条关联规则是否有趣可以采用客观兴趣度和主观兴趣度两类评价标准[2]: 客观兴趣度指那些由数据本身的属性决定的因素, 这些因素一般都是可以通过数学的方法定量计算的; 主观兴趣度度量在评价规则的有趣性时将用户的需求和系统更加紧密的结合起来, 需要来自领域专家的大量先验信息. 本文主要针对客观兴趣度进行探讨.

① 收稿时间:2013-03-13;收到修改稿时间:2013-04-25

2 关联规则基本评价指标

2.1 支持度和置信度

支持度 Support(A→B) 即项集 A 和项集 B 的并集 A∪B 在所有事务 D 中出现的概率, 取值范围为[0, 1]. 如果 A 和 B 在交易数据库中都没有出现, 则 Support(A→B)=0; 如果 A 和 B 在交易数据库中的每个交易事务中都出现, 则 Support(A→B)=1. 支持度的度量反映了关联规则是否具有普遍性, 支持度高说明这条规则可能适用于数据集中的大部分事务. 最小支持度即用户规定的关联规则必须满足的最小支持度, 它表示了一组物品集在统计意义上的需满足的最低程度.

置信度 Confidence(A→B)即在出现了项集 A 的事务 D 中, 项集 B 也同时出现的概率, 取值范围为[0, 1]. 如果 A 和 B 无关, 则 Confidence(A→B)=0; 如果 A 出现 B 一定出现, 反过来 A 不出现 B 也有可能出现, 则 Confidence(A→B)=1. 置信度的度量反映了关联规则的可靠性, 置信度高说明如果满足了关联规则的前件, 同时满足后件的可能性也非常大. 最小置信度即用户规定的关联规则必须满足的最小置信度, 它反应了关联规则的最低可靠度.

2.2 “支持度-置信度”架构的局限性

强关联规则即同时满足最小支持度阈值和最小置信度阈值的规则. 关联分析的任务就是找出数据集中隐藏的强规则. 尽管在生成关联规则的过程中,利用支持度和置信度进行剪枝大大减少了生成的关联规则的数量,但是不能完全依赖提高支持度和置信度的阈值来筛选出有价值的关联规则. “支持度-置信度”架构存在着一定的缺陷: 如果支持度阈值设置过低, 算法所需的计算量和内存需求将增加, 同时会生成过多的关联规则, 其中有些关联规则可能是虚假的规则; 如果支持度阈值设置过高, 就会有丢失一些重要规则的风险, 从而有可能丢失用户观点来看是有意义的规则问题. 比如商场奢侈品的购买记录, 虽然只占很小的比例, 也就是支持度比较低, 但由于奢侈品的利润高, 它的购买模式对于商场来说非常重要. 如果支持度阈值和置信度阈值都很高,则产生的关联规则往往是显而易见, 早已经被掌握的知识. 而置信度的缺陷在于该度量只考虑了 A 与 AB 的关系, 忽略了规则后件中项集的支持度, 有时也不能正确反映前件和后件之间的关联.

下面我们来看一个例子: 假设对 5000 名商场购物的顾客一次购买烟酒的情况进行统计, 如表 1 所示. 设支持度阈值为 40%, 置信度阈值为 60%.

表 1　某超市烟酒销售数据

	购买酒	不购买酒	合计
购买烟	2000	1750	3750
不购买烟	1000	250	1250
	3000	2000	5000

采用支持度、置信度对数据进行关联分析, 关联规则购买酒→购买烟的支持度 Support=2000/5000=40%, 置信度 Confidence=2000/3000=67%, 显然购买酒→购买烟这条规则是强关联规则, 表明购买酒的顾客通常也会购买烟. 但是在所有的顾客中购买烟的顾客比例为 3750/5000=75%,要大于 67%. 这说明一个顾客如果购买酒, 那么他购买烟的可能性就从 75%降低到 67%, 而且不购买酒→购买烟的可能性为 1750/2000=87.5%, 因此, 如果将这条关联规则提供给决策者, 就会产生误导, 因为购买酒反而会抑制购买烟, 烟和酒之间并不是一个令决策者感兴趣的关联规则.

3 基于相关性的兴趣度

在分析相关性兴趣度之前, 先给出如下定义:

定义 1: P(A)表示事务中 A 发生的概率, P(AB)表示事务中 A 和 B 同时发生的概率,若 P(AB)= P(A)P(B), 则 A 和 B 相互独立; 若 P(AB)≠P(A)P(B), 则 A 和 B 不相互独立. 此定义可推广到多个项目. 如若 P(ABC)=P(A)P(B)P(C),则 ABC 相互独立; 如若 P(ABC)≠P(A)P(B)P(C),则 A、B、C 不相互独立.

定义 2: 在 n 次事务中,count(A)表示 A 在 n 次事务中出现的次数,则 $P(A)=\frac{count(A)}{n}$

3.1 兴趣度的定义

由上面的例子分析可以看出, 强关联规则不一定是有趣的, 不仅可能没有实用价值, 甚至有可能是有误导性的和错误的. 为了避免生成“错觉”的关联规则, 在支持度-置信度架构的基础上增加了基于相关性的兴趣度来评价一个规则的优劣.

$$RI(A\rightarrow B)=\frac{p(AB)}{p(A)p(B)}\ [3] \qquad (1)$$

RI 表示规则 AB 的兴趣度, 用于评估 A 的出现“提升”B 的程度,既考虑了 p(A), 也考虑了 p(B). 取值范围是[0, +∞), 描述正相关的区域和描述负相关的区间是

不对称的. 如果 RI>1,则 A 的出现和 B 的出现是正相关的, 取值范围是(1, +∞),意味着 A 的出现可以带动 B 的出现, RI 的值越大, 则 A 对 B 的带动越大, 这是商场销售分析中需要的; 如果 RI<1,则 A 的出现和 B 的出现是负相关的,取值范围是(0, 1), 即 A 的出现降低了 B 出现的可能, 这是商场销售分析中不需要的; 如果 RI=1, 说明 A 的出现和 B 的出现是相互独立的, 没有相关性.

以表 1 中的数据为例采用式(1)计算购买酒→购买烟的兴趣度: P(购买烟和酒)=2000/5000=0.4, P(购买酒)=3000/5000=0.6, P(购 买 烟)=3750/5000=0.75, RI=P(购 买 烟 和 酒)/(P(购 买 酒)P(购 买 烟))=0.4/(0.6*0.75)=0.89, 由于购买酒→购买烟的 RI 小于 1, 所以前后件之间存在负相关关系, 即购买酒不但不会提升购买烟的人数, 反而会减少.

$$RI(A\rightarrow B)=P(AB)-P(A)P(B)^{[4]} \quad (2)$$

RI 表示规则 AB 的兴趣度, 取值范围是(-0.25, +0.25), 如果 RI<0, 则 A 的出现和 B 的出现是负相关的; 如果 RI=0, 则 A 的出现和 B 的出现是相互独立的, 没有相关性; 如果 RI>0,则 A 的出现和 B 的出现是正相关的. 计算 RI=0.4-0.6*0.75=-0.05<0, 同样可以得知购买酒→购买烟是负相关关系.

3.2 兴趣度的特性

通过式(2)我们可以分析得出如下数学特性[5]: 当 AB、A、B 的支持度为 0.5 时, A、B 的正相关度最大为 0.25; 当 AB 的支持度为 0, A、B 的支持度为 0.5 时, A、B 的负相关度最大为-0.25; A、B 的支持度很大或很小时两者的相关度都很小; A、B 正相关的充要条件是 B、A 负相关; A、B 的相关度与 B、A 的相关度是一对相反数. 所以式(2)用来度量规则的兴趣度比式(1)更合理. 采用式(2)计算规则的兴趣度, 在获取有趣规则时, 不用同时采用支持度、置信度和兴趣度 3 种阈值加以限制, 只要采用兴趣度就可以了. 给定一个兴趣度阈值后, 既不会得到负相关的规则, 也不会得到矛盾的规则.

3.3 趣度的局限性

式(1)可以评估 A 的出现是否能够促进 B 的出现, 但也有不足之处, 问题是其对阵性, RI(A→B)和 RI (B→A)值都是 $\frac{p(AB)}{p(A)p(B)}$ 因此判断不出 A→B 和 B→A 哪个规则更有趣. 很显然这不一定符合实际情况, 因为购买 A 的人可能倾向于购买 B, 并不意味着购买 B 的人就一定倾向于购买 A. 就像人们买肉的时候可能也会买蔬菜, 但是买蔬菜的时候不一定会买肉一样. 而式(2)默认支持度是 0.5, 忽略了不同领域用户的主观感觉,不同的用户可能对不同类型的规则感兴趣.

4 总结

关联规则挖掘产生了成千上万的规则, 怎样的衡量标准才是有价值的, 能否产生所有有价值的模式, 能否只产生有价值的模式等关联规则评价 3 个层次的问题. 针对兴趣度的关联规则评价方法研究已经成为一个重要的研究方向. 本文首先分析了“支持度-置信度”的度量方法, 通过实例指出此架构的局限性, 提出了基于相关性的兴趣度的评价指标 ps 公式, 针对 PS 公式的数学特性指出了它的优点与不足, 为关联规则评价体系的改进奠定了基础.

参考文献

1 窦祥国,胡学刚.关联规则的评价方法研究.安徽技术师范学院学报,2005,(4):44–47.
2 蒋盛益,李霞,郑琪.数据挖掘原理与实践.北京:电子工业出版社,2011.
3 谢文阁,梅红岩,等.基于兴趣度的关联规则在选课分析中的应用.内蒙古大学学报,2009,(3):199–202.
4 张铁军.关联规则挖掘的相关问题研究[硕士学位论文].西安:西安科技大学,2009.
5 梅志芳,王建.关联规则兴趣度问题研究.计算机工程,2010,(1):38–42.