**浅谈关联规则**

作者：张青松

目录

[关联规则基本理论 2](#_Toc420155666)

[1.1. 关联规则基本概念 2](#_Toc420155667)

[1.2. 关联分析 3](#_Toc420155668)

# 关联规则基本理论

关联规则主要用于从数据集中发现数据项（item）之间的联系，又称作购物篮分析（market basket analysis），最著名的应用实例就是啤酒尿布推销策略。20世纪90年代，美国沃尔玛曹氏管理人员共同分析销售数据发现，啤酒与尿布两个毫不相关的商品经常会被同时购买。进一步分析发现，在有婴儿的家庭中，大多数是母亲在家照顾婴儿，父亲去超市买尿布，在购买尿布的同时，顺便购买自己喜欢的啤酒，这就导致了啤酒与尿布的这种奇怪的组合。根据这一现象，超市将啤酒与尿布摆放在临近的位置，达到了双向营销的目的。

## 关联规则基本概念

为说明关联规则的一些概念，通过举例的方法进行。

如果将商场销售的所有商品种类设为一个集合，每种商品（item）为一个二元的分类变量，其取值为True或者False，用来描述时候被某个客户购买。每个客户的购物车就是一个boolean值的向量。通过分析这些向量，就可以得到哪些商品往往是被同时购买的。下表显示了四种商品在七个购物车中的购买情况

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Item1 | Item2 | Item3 | Item4 |
| Shopping1 | 1 | 0 | 1 | 0 |
| Shopping2 | 0 | 0 | 1 | 1 |
| Shopping3 | 1 | 1 | 1 | 1 |
| Shopping4 | 1 | 1 | 0 | 0 |
| Shopping5 | 0 | 0 | 1 | 0 |
| Shopping6 | 1 | 0 | 1 | 1 |
| Shopping7 | 0 | 1 | 1 | 1 |

* 规则：规则形如“如果。。。那么。。。”的结构，前者表条件，后者表结果。
* 所有项集（items）：所有商品的集合。
* 记录（Transaction）：每一个购物差为一个记录。
* 资料库（Transaction Database）：所有记录的集合。
* 项集（itemset）：同时被购买的商品的集合。
* 信任度（Confidence）：用来度量一个规则是否够好的度量之一。观察上表，在同一个购物车中，当item4被购买了，那么一定也购买了item3，那么item4对于item3的信任度为100%。将item4记为A，Item3记为B：

$$confidence\left(item4=>item3\right)=P\left(A\right)=\frac{P\left(AB\right)}{P\left(A\right)}=\frac{4}{4}=100\%$$

* 支持度（support）：衡量规则的另一个标准。从上表可以看出，虽然item3和item4常常同时购买，但是在整个集合中购买item3和item4的情况相对较少，因此单由信任度得到的结论不足以说明规则。支持度的公式：

$$support\left(item4=>item3\right)=P(AB)$$

* 提升度（lift）：从上表可以看出item3自身的支持度support(item3)=P(B)=6/7=85.71%，表示不做任何推荐时，客户购买item3的概率为85.71%。而用户购买了item4后，购买item3的概率是100%，购买item3的概率提高了100/85.71=1.667。这个概率之比称作提升度。提升度用于关联规则是否具有可用性,如果提升度大于1表示该关联规则可用，反之，使用了关联规则还不如未使用，表示该关联规则无效。公式如下：

$$lift\left(item4=>item3\right)=\frac{confidence\left(A=>B\right)}{support\left(B\right)}$$

## 关联分析

如果设置取值范围为[0%,100%]的最下支持阈值min\_sup和最小信任阈值min\_confid，则关联分析的实质就是在项集中找到符合以下规范的关联规则A=>B，并根据提升度找出可用的关联规则。

$$support\left(A=>B\right)>min\\_sup$$

$$confidence\left(A=>B\right)>min\\_confid$$

$$lift\left(A=>B\right)>1$$