APR 算法知识拓展

在本关课程中我们学习了 APR 算法相关的内容，这其中其实涉及到了一些统计学的知识以及关联分析的概念。

有的同学可能看完之后还是不太明白，所以这里特意给大家补充一些相关知识，帮助大家对于为什么要用支持度、置信度以及提升度去衡量一条强关联规则的有效与否有更深入的理解。

首先是一些统计学的基础知识，相信很多同学在学生时代都接触过概率的相关内容，它的定义是这样的：概率、也被称为“或然率”，是介于 0~1 之间的一个值，它是对某事件发生的可能性大小的一种数值度量。

我们一般用 P(A) 来表示事件 A 发生的概率。

举个例子来说，假定在相同条件下，重复进行 n 次试验，事件 A 发生了 m 次，则事件 A 发生的概率可表示为：

$$P(A) = \frac{事件A发生的次数}{重复试验次数} = \frac{m}{n}$$

其中概率值越接近于 0，说明事件发生的可能性就越低；概率值越接近 1，说明事件越有可能发生。

它的作用是，当面临选择问题时，概率会用数值表达哪种可能性最大，帮助我们做出最好的选择。虽然概率并不会准确地告诉我们将会发生什么，但概率能够告诉我们很有可能发生什么，不太可能发生什么。

在概率的分支里，还有一种条件概率。它表示事件 A 在另外一个事件 B 已经发生条件下的发生概率。条件概率表示为： P(A|B) ，意思是：给定事件 B 的条件下，事件 A 发生的概率。公式表示为：

$$P(A|B) = \frac{P(AB)}{P(B)}$$

关于它们的含义，P(A|B)：A 表示事件 A，B 表示事件 B，P(A|B)代表事件 B 已经发生的情况下，事件 A 发生的概率。
P(AB)：表示事件 A 和事件 B 同时发生的概率。
P(B)：表示事件 B 发生的概率。

条件概率的作用是在关联事件中，计算事件发生后另一个事件发生的概率。而我们本关所学的置信度实际上就是一个条件概率。

课程中对于 {X} 的支持度定义为：{X} 在事务中出现的次数 / 事务总数。那么根据概率的定义，支持度的公式就可以写为：

$$support\{X\} = P(X)$$

即 X 在所有事务中出现的概率。

而对于 {X} –> {Y} 的置信度定义为：{X, Y} 的支持度 / {X} 的支持度。那么置信度的公式就可以写成：

$$confidence\{X\} \rightarrow \{Y\} = P(Y|X) = \frac{P(XY)}{P(X)}$$

即 X 发生的情况下，Y 出现的概率。

最后，对于 {X} –> {Y} 的提升度课程里面定义为：{X} –> {Y} 的置信度 / {Y} 的支持度。那么提升度的公式可以写为：

$$lift\{X\} \rightarrow \{Y\} = \frac{P(Y|X)}{P(Y)}$$

即 X 发生的情况下，Y 发生的概率和 Y 发生的概率的比值。

搞清楚了这些，我们就可以来思考一下支持度、置信度、提升度的意义了。

置信度（条件概率）的本质是反应两个事件同时发生的概率大小。置信度（条件概率）的值越大，代表事件 A 的发生对事件 B 的发生影响越大，那么事件 B 越有可能发生。

但是置信度只是衡量一个关联规则是否有效的一个指标，真正要确定一条关联规则是否有效，还需要考虑以下两个指标：

支持度：P(AB)
提升度：P(B|A) / P(B)

于是我们接着来看支持度的意义，支持度是用来衡量某个项（某些项）在所有事务中出现的概率，概率越高，说明客户对于某个项（某些项）的购买意愿越强烈。

这样一来就会存在一个问题：如果关联规则 {X} –> {Y} 的置信度 P(Y|X) 很高，但是 {X, Y} 的支持度 P(XY) 很低，就意味着用户对于同时购买商品 X 和 Y 的意愿非常低，即便之后将 X 和 Y 进行捆绑销售，销量 X 或者 Y 的销量很有可能依旧很低，难以被提升。

所以不难看出支持度是衡量关联规则是否有效的一个重要指标。反映的是用户对于关联规则中出现的商品的购买意愿强弱情况，也反映了这些商品的销量情况。

关联规则本质上是想要把某些商品组合在一起来提高这些商品的销量，但如果这些商品本身的销量就特别低，那即便放在一起了，也很难达到促进的效果。

最后我们再来看看提升度的意义。提升度是用来衡量关联规则 {X} –> {Y} 中，X 的出现对 Y 的影响程度的，也就是研究 X 的出现是促进了 Y 的出现还是抑制了 Y 的出现。

这样同样会有一个问题。

我们都知道置信度是用来衡量 X 的出现对 Y 的影响程度的，但是会有一个缺陷：如果 Y 本身出现的次数就特别多，以至于无论 X 出不出现，Y 出现的概率都特别高，或者说，X 如果不出现，Y 出现的概率会比 X 出现时更高。

那么即便 P(Y|X) 满足了我们的要求，却也无法被当作一条有效的关联规则。

举个例子来说，提升度反映的是关联规则 {X} –> {Y} 中，X 对 Y 的促进或抑制作用。

那么如果 {X} –> {Y} 的提升度大于 1，也就是 P(Y|X) > P(Y) ，这意味着 X 出现的情况下，Y 出现的概率要比 Y 整体出现的概率要高。那么就可以看作是 X 对 Y 有促进作用，关联规则有效。

如果 {X} –> {Y} 的提升度等于 1，也就是 P(Y|X) = P(Y)，这就意味着 X 出现的情况下，Y 出现的概率和 Y 整体出现的概率一样高。那么就可以看作是 X 对 Y 既没有促进作用，也没有抑制作用，关联规则无效。

最后，如果 {X} –> {Y} 的提升度小于 1，也就是 P(Y|X) < P(Y) ，这意味着 X 出现的情况下，Y 出现的概率要比 Y 整体出现的概率要低。那么就可以看作是 X 对 Y 有抑制作用，关联规则无效。

所以提升度同样也是衡量关联规则是否有效的一个重要指标。

好啦，今天的分享就先到这里啦，刚才所讲的这些关于概率的知识光看一遍可能还无法理解，所以我特地准备了一份资料给大家，里面通过一些具体的例子把我刚才讲的知识融入了进去，希望可以对大家理解关联分析有所帮助：
https://docs.forchange.cn/docs/e1Az4V6G8WfbR2qW/ 《关联分析教辅资料》
