RFM 模型学习笔记

李向阳　d1142845997@gmail.com

目录

1 引入

这一次, 我们来谈谈 RFM 模型.

我们知道机器学习中的算法大致可归为回归、分类、聚类、预测等等, 不过数据挖掘中的很多算法似乎难以归类. 勉强对应的说, RFM 模型是一种聚类的方法, 即将用户分为几类, 当然, RFM 模型不仅可用于聚类, 它的应用也很多.

RFM 模型中的特征变量是固定的, 就是指如下 3 个变量

- Recency: 最近一次消费
- Frequency: 消费频率
- Monetary: 消费金额

也就是依照这 3 个变量, 我们来挖掘用户的信息并加以利用, 这样的模型便称为 RFM 模型.

2 RFM 模型用于聚类

我们来看最典型的例子, 即将 RFM 模型用于聚类.

2.1 数据集

以超市购物为例, 我们的样本数据集的一部分可能如下表1, 当然, 这个完整的表是用 R 语言人工生成的.

其中的 ID 就是指顾客 ID, 相应的 Date 和 Amount 就是指该顾客当日的消费金额. 那么如何得到我们需要的 3 个变量呢?

其实就是选定一个当前时间, 比如 1998-07-01, 这样就可以计算变量 Recency, 也即每个顾客的最近一次消费了 (按间隔的天数计算). 而计算变量 Frequency, 只要统计顾客购买了几次就可以了. 至于变量 Monetary, 我们一般用顾客平均每次的购买额来代替, 即用总的消费额除以消费次数. 经过如此计算后, 可以得到我们的数据集, 如下表2.

这样就得到 RFM 模型的数据集了.

2.2 RFM 模型聚类

得到数据集后, 我们需要对数据进行标准化处理, 并称为每个变量的得分. 这里有两种方式, 一种是把 R, F, M 都对应成评级 1 到 5 分, 另外一种

表 1: 原始数据集

ID	Date	Amount
4	1997-01-01	29.33
4	1997-01-18	29.73
4	1997-08-02	14.96
4	1997-12-12	26.48
21	1997-01-01	63.34
21	1997-01-13	11.77
50	1997-01-01	6.79
71	1997-01-01	13.97
86	1997-01-01	23.94
111	1997-01-01	35.99
111	1997-01-11	32.99
111	1997-03-15	77.96
111	1997-06-23	91.92
111	1997-07-22	47.08
111	1997-07-26	71.96
111	1998-05-10	72.99
111	1998-06-20	55.47
112	1997-01-01	11.77
112	1997-02-05	11.77
113	1997-01-01	32.91
113	1998-03-04	15.27
113	1998-03-07	11.49
114	1997-01-01	16.36
114	1997-05-01	28.13

是普通的标准化处理 (当然, 数据标准化有很多种方法). 以最大最小标准化为例, 我们得到如下表3.

数据标准化后, 我们需要计算每个顾客的加权总得分, 这里的权值怎么确定呢? 网上主要也有两种方式, 一种是把 R, F, M 的权重分别赋为 100, 10, 和 1, 而且这种处理方法通常和标准化时评级成 1 到 5 分相结合, 这样, 若一个顾客得了 542 分, 说明他的 Recency 得了 5 分, Frequency 得了 4 分, Monetary 得了 2 分. 另外一种是采用层次分析法得到这三个变量的权重 (我会在下面的补充中给一个简单例子), 而且这种处理方法通常和标准化

表 2: 数据集

ID	Recency	Frequency	Monetary
4	201	4	25.125
18	543	1	14.96
21	534	2	37.555
50	546	1	6.79
60	515	1	21.75
71	546	1	13.97
86	546	1	23.94
111	11	16	69.19
112	511	2	11.77
113	116	3	19.89
114	140	5	24.986
131	546	1	30.32
133	232	7	28.453

表 3: 标准化数据集

ID	Recency	Frequency	Monetary
4	0.633	0.055	0.05
18	0.006	0	0.03
21	0.022	0.018	0.074
50	0	0	0.013
60	0.057	0	0.043
71	0	0	0.028
86	0	0	0.047
111	0.982	0.273	0.136
112	0.064	0.018	0.023
113	0.789	0.036	0.039
114	0.745	0.073	0.049

时普通的标准化相结合.

除此之外, 也可以采用熵值法来确定每个指标的权重, 然后算加权总得分, 比如我们最终算得 R、F、M 的权重分别为 $0.3, 0.1, 0.6$, 那么可以算出每个用户的总得分如下表4(用得分矩阵乘以权重向量即可).

顾客的总得分越高, 说明这个顾客对我们来说越重要. 除此之外, 我们

表 4: 用户总得分

ID	Recency	Frequency	Monetary	TotalScore
4	0.633	0.055	0.05	0.2254
18	0.006	0	0.03	0.0198
21	0.022	0.018	0.074	0.0528
50	0	0	0.013	0.0078
60	0.057	0	0.043	0.0429
71	0	0	0.028	0.0168
86	0	0	0.047	0.0282
111	0.982	0.273	0.136	0.4035
112	0.064	0.018	0.023	0.0348
113	0.789	0.036	0.039	0.2637
114	0.745	0.073	0.049	0.2602

还可以计算出总的 Recency, Frequency, Monetary 的均值, 然后对每个顾客依照这三个变量的得分对顾客进行分类. 依照均值比较的方法可以分出 8 类 (因为每个分值要么高于均值, 要么低于均值), 如下表5

表 5: 分类表

Rank	Recency	Frequency	Monetary	客户类型
1	高于均值	高于均值	高于均值	最有价值客户
2	高于均值	低于均值	高于均值	重要发展客户
3	低于均值	高于均值	高于均值	重要保持客户
4	低于均值	低于均值	高于均值	重要挽留客户
5	高于均值	高于均值	低于均值	一般价值客户
6	高于均值	低于均值	低于均值	一般发展客户
7	低于均值	高于均值	低于均值	一般保持客户
8	低于均值	低于均值	低于均值	一般挽留客户

不过有时人们并不想这样简单的分为 8 类, 毕竟每个行业是不太一样的. 因此, 对顾客分类还有很多其他方法, 基本上归为两类. 一类方法是 Nested, 也就是对这 3 个变量逐个分类, 比如先依照变量 Recency 分类, 把 Recency 划分为 0-120, 120-240, 240-450, 450-500 和 500 以上. 接下来对每个大类再按照剩下的两个变量进行分类, 切割的区间也可适当划分. 另一类方法是 Independent, 即独立的在每个维度上进行分类.

3 关于 RFM 模型的补充

3.1 层次分析法确定权重

前面提到, 变量权重的确定可以采用层次分析法. 关于层次分析法, 可参见韩中庚的《数学建模方法及应用》. 其实就是求评价矩阵最大特征值对应的特征向量, 然后再归一化即可. 近似计算时, 可以采用和法 (可参考其它文献), 以 http://wiki.mbalib.com/wiki/RFM%E6%A8%A1%E5%9E%8B 的评价矩阵为例, 代码如下

```
b <- matrix(c(1, 0.71, 0.46, 1.41, 1, 0.85, 2.18, 1.18, 1),
    ncol = 3, nrow = 3, byrow = T)

f <- function(x) {
    x / sum(x)
}

b <- apply(b, 2, f)

x <- apply(b, 1, sum)
w <- f(x)
```

以上计算 w 的结果即为 $(0.221, 0.341, 0.439)^T$.

3.2 熵值法确定权重

我们知道熵可以衡量一组数据分布的离散情况. 那么如何将它运用到变量权重的确定上呢?

我们以 F 和 M 值为例, 显然 F 的值分布的估计是比较均匀的, 而 M 的值一般差异较大, 因此计算出来的熵肯定也会更大, 但是我们看的是指标的重要程度, 因此可以用 1 减去熵值作为权重. 具体应用到 RFM 模型中, 我们来说明一下计算方法.

首先对数据进行标准化处理. 本模型的样本数据矩阵为 $X = (x_{ij})_{m\times n}$, 其中假设有 m 个样本, 有 $n = 3$ 个变量, 示例可见表2. 比如进行最大最小标准化, 由于 F, M 是正向指标 (即越大表现越好), 因此标准化公式为

$$x_{ij}^* = \frac{x_{ij} - \min\{x_j\}}{\max\{x_j\} - \max\{x_j\}}$$

而 F 是负向指标 (即越小表现越好), 因此标准化公式为

$$x_{ij}^* = \frac{\max\{x_j\} - x_{ij}}{\max\{x_j\} - \max\{x_j\}}$$

接着计算

$$y_{ij} = \frac{x_{ij}^*}{\sum_{i=1}^{m} x_{ij}^*}$$

然后计算变量的信息熵

$$e_j = \frac{1}{\ln m} \sum_{i=1}^{m} y_{ij} \ln y_{ij}$$

注意当 $y_{ij} = 0$ 时, 可以规定 $\ln y_{ij} = 0$, 这个我们在讲决策树计算熵时也遇到过.

最后计算各个变量的信息熵冗余度并得到权重

$$d_j = 1 - e_j, w_j = \frac{d_j}{\sum_{j=1}^{n} d_j}$$

4 总结

4.1 参考资料

(1) MBA 智库百科: http://wiki.mbalib.com/wiki/RFM%E6%A8%A1%E5%9E%8B

(2) 博客: http://ljy.logdown.com/posts/2014/12/27/rfm-analysis-using-r, 用 R 语言生成数据并进行了简单分析.

(3) 博客: http://www.dataapple.net/?p=84, 比较完整的介绍, 基于 R 语言, 也有数据集.

(4) 博客: http://www.marketingdistillery.com/2014/11/02/rfm-customer-segmentation-in-r- 不仅有 R, 还有 Python 和 Spark 的数据处理.

参考文献

[1] 李荣华. 偏微分方程数值解法. 高等教育出版社 (2010)

[2] Zhilin Li,Zhonghua Qiao,Tao Tang.*Numerical Solutions of Partial Differential Equations-An Introduction to Finite Difference and Finite Element Methods.*(2011)

[3] 孙志忠. 偏微分方程数值解法. 科学出版社 (2011)

[4] 陆金甫关治. 偏微分方程数值解法. 清华大学出版社 (2004)

附录