基于 TOPSIS 法的信用卡客户违约风险度量

中央财经大学　　司徒雪颖

一、研究目的

信用卡客户的违约风险向来是银行等金融机构关注的重点之一，然而如何衡量客户违约风险往往是信用卡业务中的重要问题。本文使用违约发生时间和违约严重程度来构建违约风险度量，并按照该度量模式对各类客户的违约风险从高到低进行排序，可为金融机构提供一些参考。

二、数据预处理及数据介绍

数据集为信用卡违约数据，包括人口学变量（性别、年龄、受教育程度、职业）和违约发生时间和违约严重程度。由于其他变量如年龄，有诸多与事实不相符之处，因此直接删除，只使用违约发生时间和违约严重程度来构建违约风险度量，并按照该度量模式对各类客户进行排序。

1. 违约发生时间的处理

仅选取违约发生时间在 2016-1-1 到 2017-12-31 之间的样本,其他样本视为异常值删除。然后对这两年的数据以月为单位设定月份序列，2016 年 1 月为第 1 个月，2016 年 2 月为第 2 个月，以此类推，2017 年 12 月为第 24 个月。违约发生时间距今越近，违约风险越高。

2. 违约严重程度的处理

原数据的违约严重程度使用“轻度”,“中度”,“重度”三个水平来描述，本文为量化违约严重程度，令“轻度”为 1,“中度”为 1,“重度”为 3。

3. 处理后的数据介绍

处理后的数据集由 65368 条数据，3 个变量构成，含缺失值的 10 行已删除。

列变量如下表 1 所示：

表 1 构建违约风险度量的指标

变量名	说明	取值范围
顾客代码	定性变量 共 12 类顾客	1~12
违约发生时间	定序变量 单位：月	从 2016-1-1 到 2017-12-31， 1 代表 2016-1， 2 代表 2016-2， ……， 14 代表 2017-12
违约严重程度	定序变量 共 3 个水平	1 代表轻度 2 代表中度 3 代表重度

三、描述统计

1.单变量描述统计

如图 1（A）所示，轻度违约的人数最多，为 52610 人，占总数 80.48%，约是中度违约人数（11643 人）的 4.5 倍，重度违约人数为 1115 人，占总数 1.71%。

如图 1（B）所示，违约时间集中在 2016.12-2017.2 这三个月，这三个月发生的违约发生数占 2016-2017 两年违约发生数的 79.58%；2016.01-2016.06 期间，每月违约发生数不超过 1000,2016.07-2016.11 期间，每月违约发生数在 1000~2000 之间；2017.02 以后，每月违约发生数除 2017.12 为 11 例外，都不超过 10 例。

如图 1（C）所示，“顾客 2”类别人数最多，为 34894 人，是人数第二多的“顾客 1”（8063 人）的 4.3 倍，“顾客 9”“顾客 6”“顾客 7”“顾客 11”人数很少，在 300~200 人之间。

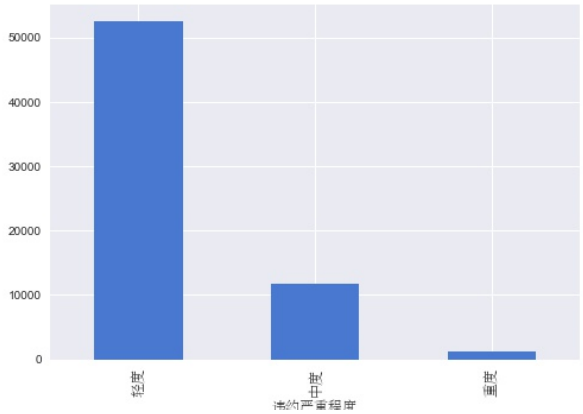

图 1（A）违约严重程度的条形图

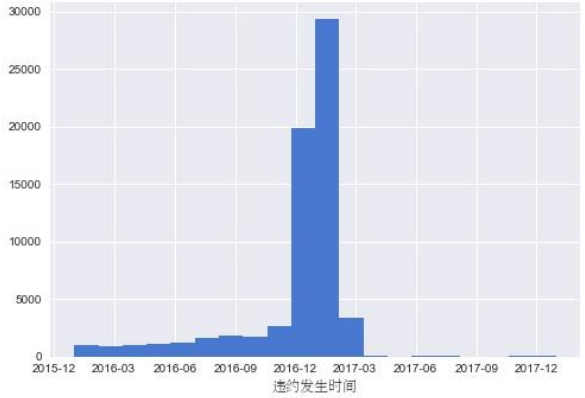

图 1（B）违约发生时间分布直方图

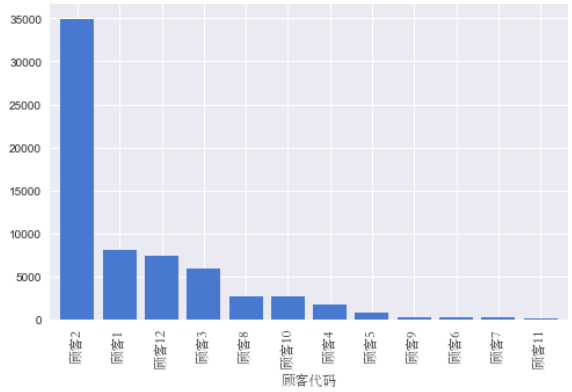

图 1 （C）顾客代码数目条形图

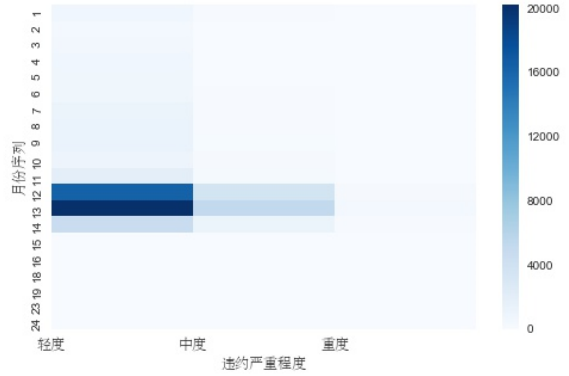

图 1（D）违约严重程度与违约发生时间的热图

2.双变量交互描述统计

（1）违约严重程度与违约发生时间

如图 1（D）所示，绝大多数的违约的发生时间集中在 2016.12 和 2017.01，违约程度为轻度居多。

（2）顾客代码与违约发生时间

图 2 同样表明，违约发生时间集中在第 12、13、14 个月（即 2016.12-2017.02），其中“顾客 6”、“顾客 9”、“顾客 11”，“顾客 4”，“顾客 5”在第 15 个月（2017.03）后无违约人数。而“顾客 11”在 2016.12-2017.02 违约的人数仅仅占“顾客 11”总数的 41.5%，相比于整体和其他顾客类别，“顾客 11”的违约时间分布较为分散，且违约时间均距今较长。从违约时间上看，“顾客 11”将会是违约风险最低的一组。

图 2 顾客代码与月份序列的小提琴图

（3）顾客代码与违约严重程度

由图 3 可以看出，“顾客 4”中轻度违约的人数占比最小（58.08%），且重度违约人数占比最大（8.67%）的一类，而“顾客 6”和“顾客 11” 轻度违约的人数占比最大，分别为 92.5%、94%，。结合前述，轻度违约占总数 80.48%，重度违约占总数 1.71%，可知从违约严重程度上看，“顾客 6”将是违约风险最低的一组。

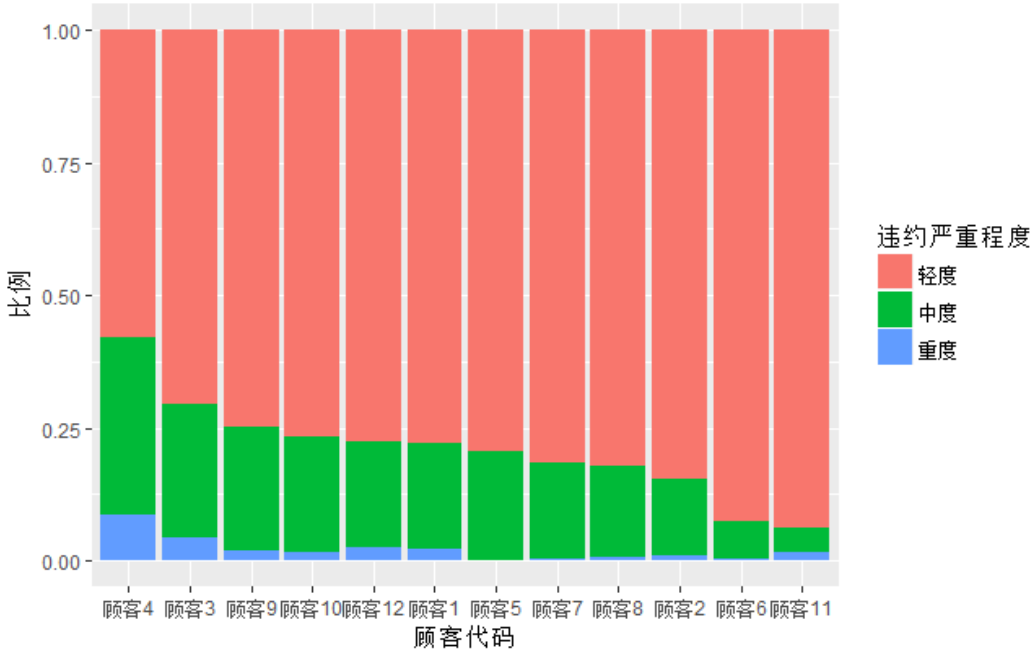

图 3 顾客代码与违约严重程度的堆积柱状图

四、违约程度量化方法

1.按顾客代码分组汇总的数据透视表

如表 2 所示，按顾客代码分组，求出违约发生时间和违约严重程度的均值和方差。可以看出，“顾客 11”在违约发生时间和违约严重程度上均值最小（分别为 9.72 和 1.08），因此，它一定是违约风险最小的一类。而违约发生时间均值最大的是“顾客 3”，违约严重程度均值最大的是“顾客 4”，因此这两类顾客可能是违约风险最大的两组。

表 2 按顾客代码分组汇总的数据透视表

顾客代码	违约发生时间		违约严重程度		
	均值	均值（标准化）	均值	均值（标准化）	计数
顾客 1	11.65	0.39	1.24	0.11	8063
顾客 10	11.59	0.29	1.25	0.17	2657
顾客 11	9.72	-2.76	1.08	-1.34	200
顾客 12	11.25	-0.25	1.25	0.16	7372
顾客 2	11.57	0.27	1.16	-0.58	34894
顾客 3	12.19	1.27	1.34	0.95	5975
顾客 4	11.50	0.16	1.51	2.40	1775
顾客 5	11.55	0.24	1.21	-0.19	863
顾客 6	11.93	0.86	1.08	-1.30	267
顾客 7	11.34	-0.19	1.19	-0.35	234
顾客 8	11.63	0.36	1.18	-0.40	2736
顾客 9	11.02	-0.64	1.27	0.36	332

把表 2 的两组均值标准化后，绘制违约严重程度与违约发生时间的散点图，X 轴表示违约严重程度，Y 轴表示违约发生时间。可以看出顾客 11 毫无悬念地成为违约风险最低的一类顾客，而顾客 3 或者顾客 4 将会是违约风险最高的两类顾客，剩下的顾客 1,10,8,5,2,7,12，9 违约风险将会比较接近，对于顾客 6，如果应用一个重点关注违约发生时间的风险衡量方法里，顾客 6 的违约风险将很高，反之则很低。

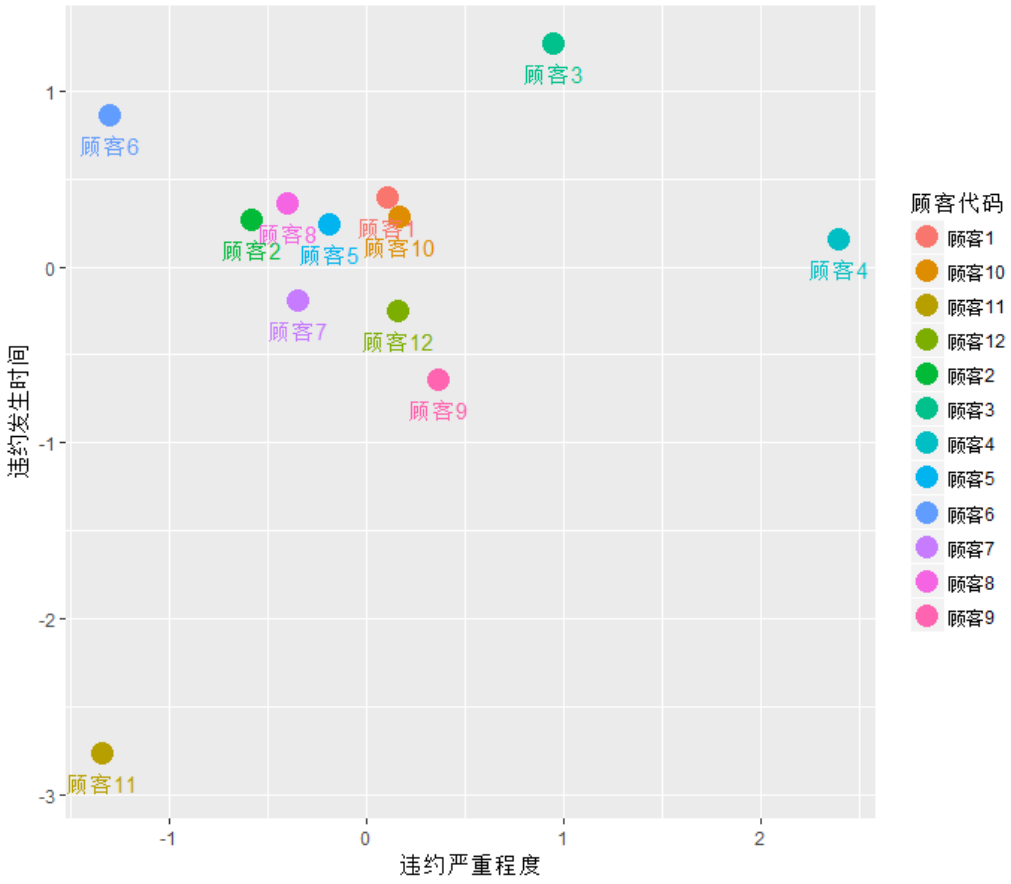

图 4 违约严重程度与违约发生时间散点图（标准化后）

2.指标综合

（1）主观赋权法

指标 1：违约严重程度均值，指标 2：违约发生时间均值。

$$\text{Score} = \alpha * \text{指标 } 1 + (1-\alpha) * \text{指标 } 2 \qquad \alpha \in (0,1)$$

令 α 等于 0.5，得到如下排名：

表 3 主观赋权法排名

顾客代码	Score	排名
顾客 4	1.27	1
顾客 3	1.11	2
顾客 1	0.25	3
顾客 10	0.23	4
顾客 5	0.02	5
顾客 8	-0.02	6
顾客 12	-0.05	7
顾客 9	-0.14	8

顾客 2	-0.16	9
顾客 6	-0.23	10
顾客 7	-0.23	11
顾客 11	-2.05	12

该方法简单，但主观性太强，α 的值选取缺乏依据，排名结果难以使人信服。该排序方法适合于专家给出 α 权重的情况。

（2）TOPSIS 法

TOPSIS(Technique for Order Preference by Similarity to Ideal Solution)法，直译为“逼近于理想解的排序方法”，是多指标决策问题中十分常用的一种方法。这种方法通过构造多指标问题的最优点和最差点，并以靠近理想解和远离最差点两个评价判据为基准，对各可行方案进行排序。

TOPSIS 法评价的基本步骤如下：

①设 X 为用于评价的矩阵，第一列为违约发生时间均值，第二列为违约严重程度均值。X 经过标准化后得到 Z，即以上两个均值标准化后的值。

$$X=\begin{pmatrix}11.65 & 1.24\\ 11.59 & 1.25\\ \cdots\cdots & \cdots\cdots\\ 11.02 & 1.27\end{pmatrix} \qquad Z=\begin{pmatrix}0.39 & 0.11\\ 0.29 & 0.17\\ \cdots\cdots & \cdots\cdots\\ -0.64 & 0.36\end{pmatrix}$$

②由各项指标最优值和最劣值分别构成最优值向量 Z+ 和最劣值向量 Z-。由于 Z 值越小越好，因此最优点正好是顾客 11 的样本点坐标（即 Z 两列的最小值），而最差点是顾客 4 的横坐标和顾客 3 的纵坐标组成的点（即 Z 两列的最大值）。

Z^{-}=（1.27,2.40），Z^{+}=（-2.76，-1.34）

③计算各评价单元与最优值和最劣值的距离。

$$D_i^{+}=\sum_{i=1}^{12}(Z_i-Z^{+})^2 \qquad D_i^{-}=\sum_{i=1}^{12}(Z_i-Z^{-})^2$$

④计算各评价单元与最优值的相对接近度。

$$C_i=\frac{D_i^{-}}{D_i^{-}+D_i^{+}} \qquad i=1,2,3,...,12$$

⑤按相对接近度大小排序， C_i 越大，表明第 i 个评价单元越接近最优水平。

表 4 TOPSIS 法排名

顾客代码	Score	排名
顾客 4	0.19	1
顾客 3	0.24	2
顾客 1	0.41	3
顾客 10	0.42	4
顾客 5	0.46	5
顾客 8	0.47	6
顾客 12	0.48	7
顾客 2	0.50	8
顾客 9	0.51	9
顾客 6	0.51	10
顾客 7	0.52	11
顾客 11	1.00	12

该方法同时考虑到了各个样本到最优点、最差点的距离，看似完美地对本问题进行了完美的排序，但仍然存在问题。观察到表 4 中顾客 2，顾客 9，顾客 6 的 score 均接近于 0.5，说明这三个点离最优点与最差点连线的中垂线非常近，这一点从图 4 也看得出来，因此难以评定这三个点的优劣。由此我们看到 TOPSIS 法用于综合评价还存在一定的问题，不能得到合理的排序结果。

(3)改进的 topsis 法

①取最优点（A）最差点（B）连线 AB 的延长线上点 H 使得 $|BA|=|HB|$，则 H 点称其为虚拟最差点，H 的向量值为 $Z^*=2Z^- - Z^+=(5.30,6.13)$。如此一来经过 B 点的 AH 的垂线成为 AH 的中垂线，不会出现顾客 2，顾客 9，顾客 6 无法评判的问题。

②计算 D_i^+ 和 D_i^*

$$D_i^+=\sum_{i=1}^{12}(Z_i-Z^+)^2 \qquad D_i^*=\sum_{i=1}^{12}(Z_i-Z^*)^2$$

③计算 $C_i^*=\dfrac{D_i^*}{D_i^*+D_i^+}$ $\qquad i=1,2,3,...,12$

④用 C_i^* 排序， C_i^* 越大表明第 i 个评价单元越接近最优水平。

表 5 TOPSIS 法改进版排名

顾客代码	Score	排名
顾客 4	0.57	1
顾客 3	0.59	2
顾客 1	0.69	3
顾客 10	0.70	4
顾客 6	0.71	5
顾客 8	0.72	6
顾客 5	0.72	7
顾客 2	0.73	8
顾客 12	0.74	9
顾客 7	0.75	10
顾客 9	0.75	11
顾客 11	1.00	12

如表 5 所示，显然这样改进后，顾客 2，顾客 9，顾客 6 的 score 差别拉大，可避免以上不合理结果的出现。

参考文献

[1]胡永宏. 对 TOPSIS 法用于综合评价的改进[J]. 数学的实践与认识, 2002, 32(4):000572-575.