信用卡违约客户的分类与聚类分析

中央财经大学

司徒雪颖

一、研究目的

金融改革以来，其中在消费金融商品为最活络的业务，当中又以「信用卡」的推广最为成功。发卡银行对已持卡者之信用状况的管理相当重要，由于信用风险涉及许多层面，因此近年来所被重视的数据挖掘技术就成为其重要的一项工具。若能及早发现可能产生呆帐等违约情形之持卡者，对其行为进行监控，将可有效预防违约的发生。因此，本研究主要目的在于利用商业智能与数据挖掘的技术整合，利用 UCI 中信用卡违约客户数据，使用聚类模型探索客户分类，并利用分类算法建立一套相对稳定且有效的预测模型，提供相关部门与发卡机构一个准则，以降低违约比例，进而降低信用风险。

二、数据描述

数据一共 3000 条样本，不存在缺失值，包括的变量如表 1 所示。其中聚类分析剔除 I 客户 ID 和下个月是否违约两个变量，分类分析以除客户 ID 的变量为自变量，下个月是否违约为因变量，建立分类模型。

表 1 变量描述

中文名	变量名	变量类型	备注
客户 ID	ID	定量	1~30000，对应 30000 条客户数据
信用额度	LIMIT_BAL	定量	1 万~100 万
性别	SEX	定性	男性 11888 人;女性 18112 人
学历	EDUCATION	定性	硕士 10585 人;本科 14030 人;高中 4917 人
婚姻状态	MARRIAGE	定性	已婚 13659 人;单身 15964;离异 323 人
年龄	AGE	定量	21~79 岁

付款状态 2005.09~ 2005.04	PAY_0	定序	-2 = 没有消费; -1 = 按时付款; 0 = 循环信贷; 1 = 已延迟付款 1 个月; 2 = 已延迟付款 2 个月; . . .; 8 = 已延迟付款 8 个月及以上;
	PAY_2	定序	
	PAY_3	定序	
	PAY_4	定序	
	PAY_5	定序	
	PAY_6	定序	
账单数额 2005.09~ 2005.04	BILL_AMT1	定量	单位：新台币
	BILL_AMT2	定量	
	BILL_AMT3	定量	
	BILL_AMT4	定量	
	BILL_AMT5	定量	
	BILL_AMT6	定量	
付款金额 2005.09~ 2005.04	PAY_AMT1	定量	单位：新台币
	PAY_AMT2	定量	
	PAY_AMT3	定量	
	PAY_AMT4	定量	
	PAY_AMT5	定量	
	PAY_AMT6	定量	
下个月是否违约	default payment next month	定量	0=没有违约;1=违约 23364 人没有违约; 6636 人违约

三、描述统计

首先从相关图矩阵中发现与下个月违约与否存在明显相关性的变量，再用具体的可视化方法展示它们之间的联系。经过描述统计发现，明显与违约与否相关的变量有信用额度、过去 6 个月的付款状态、过去 6 个月的付款金额。

1.相关图矩阵

从图 1 相关图矩阵可以看出，下个月是否违约可能与信用额度、付款状态、付款金额有关，可进一步探究。变量之间的相关性较强。

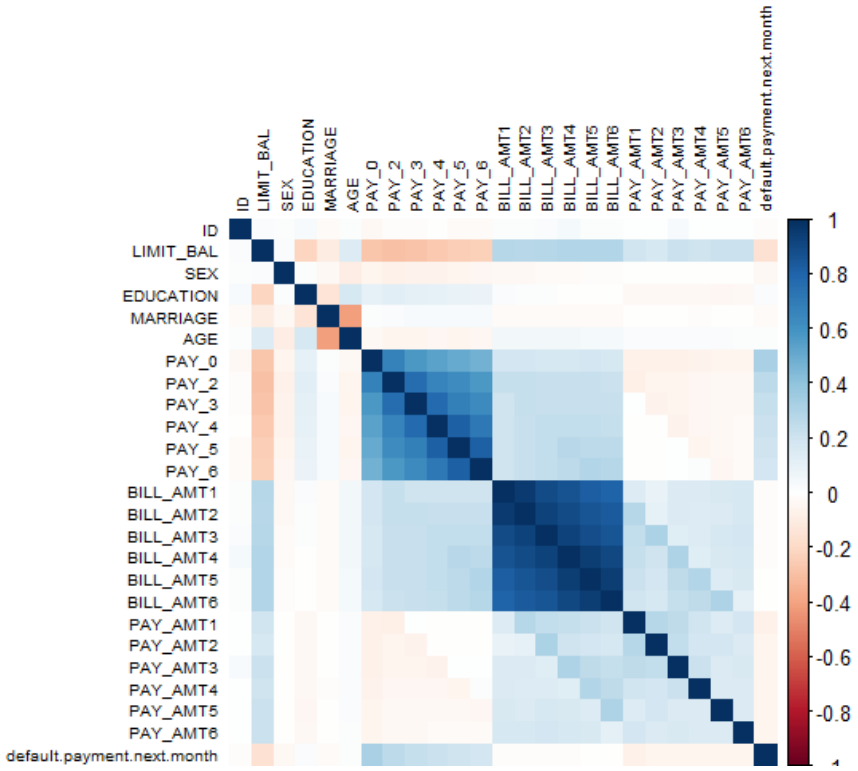

图 1 相关图矩阵

2.信用卡额度与违约与否

从图 2 贷额度的密度图可以看出，违约人群中信用额度低的较多，集中在 1 万—10 万。

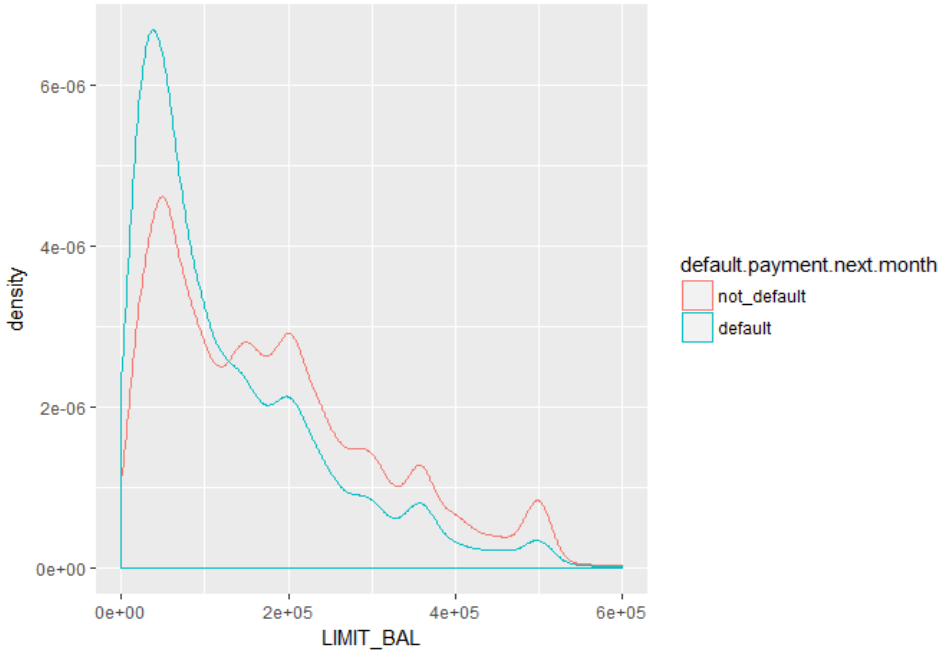

图 2 信贷额度的密度图（以是否违约分组）

3.付款状态与违约与否

图 3 堆积直方图中，红色为非违约，绿色为违约，可以看出，当 PAY_N 取值为-2,-1,0,1 时（即没有消费、循环信贷、按时付款、延迟付款一个月），违约率较低，当延迟付款大于 1 个月后，违约率明显升高至 0.5 以上。可知过去 6 个月中，出现延迟付款超过 1 一个月的客户，下个月更可能违约。

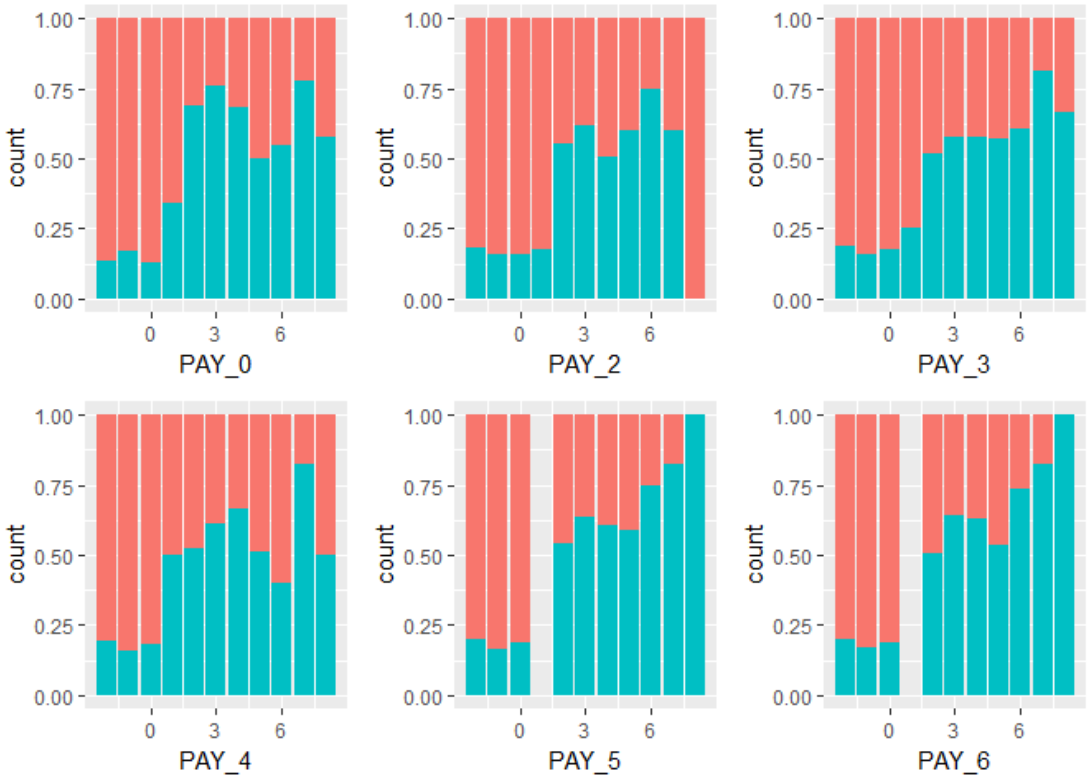

图 3　PAY 类变量与是否违约的堆积直方图

4.付款金额与违约与否

从图 4 小提琴图中可以看出，违约的客户过去 6 个月付款的金额的分布更呈现出三角形状，箱线图的 Q1 等于 0 附近，说明违约的客户付款金额多为 0 或者较少数额。从图中也可以看出，随着时间往前推移，无论是违约还是非违约客户，小提琴图都逐渐变成三角形，说明越往前看，付款少的人越多。

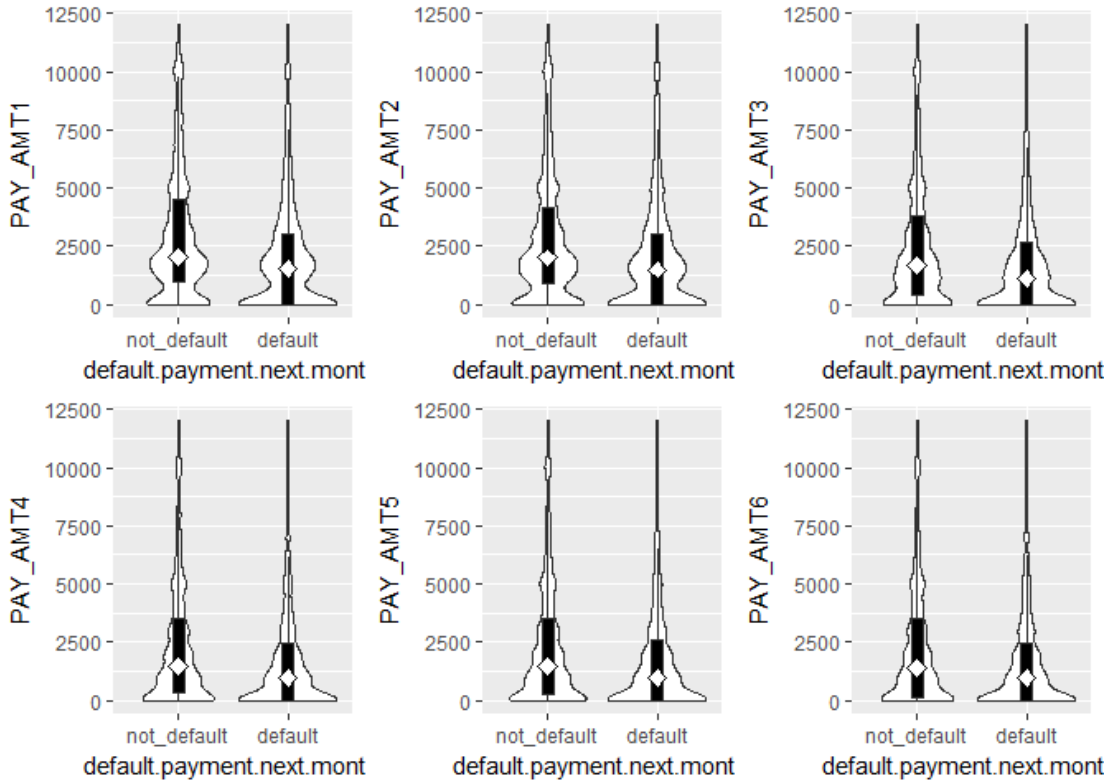

图 4 PAY_AMT 类变量与是否违约的小提琴图

四、聚类分析

由于样本数有 30000 条，考虑到计算机设备性能较弱，因此通过简单随机抽样选取 5000 条样本进行聚类。

1.PAM（partioniong around medoids）聚类

由于本数据中含有定性变量，无法使用欧式聚类，因此采用 Gower 距离进行聚类。

（1）Gower 距离定义。

Gower 距离是处理混合型数据（如同时包含连续型变量、名义型变量和顺序型变量的数据）的良好选择，首先每个类型的变量都有特殊的距离度量方法，而且该方法会将变量标准化到[0,1]之间（如数值型变量采用曼哈顿距离，分类变量先化成哑变量再利用 Dice 系数进一步计算）。接下来，利用加权线性组合的方法来计算最终的距离矩阵。

（2）PAM 算法构建聚类模型

PAM 算法的主要步骤：

1.随机选择 k 个数据点，并将其设为簇中心点

2.遍历所有样本点，并将样本点归入最近的簇中

3.对每个簇而言，找出与簇内其他点距离之和最小的点，并将其设为新的簇中心点

4.重复第 2 步，直到收敛

该算法和 K-means 算法非常相似。除了中心点的计算方法不同外，其他步骤都完全一致 。优点是简单易懂且不易受异常值所影响，缺点是算法时间复杂度为 O(n2)。

（3）聚类结果

我们将利用轮廓系数来确定最佳的聚类个数，轮廓系数是一个用于衡量聚类离散度的内部指标，该指标的取值范围是[-1,1]，其数值越大越好。通过图 5 聚类个数与轮廓系数的散点图比较不同聚类个数下轮廓系数的大小，我们可以看出当聚类个数为 10 时，聚类效果最好。考虑到分类太多可能会导致样本太过分散，因此仅尝试聚成 10 类。

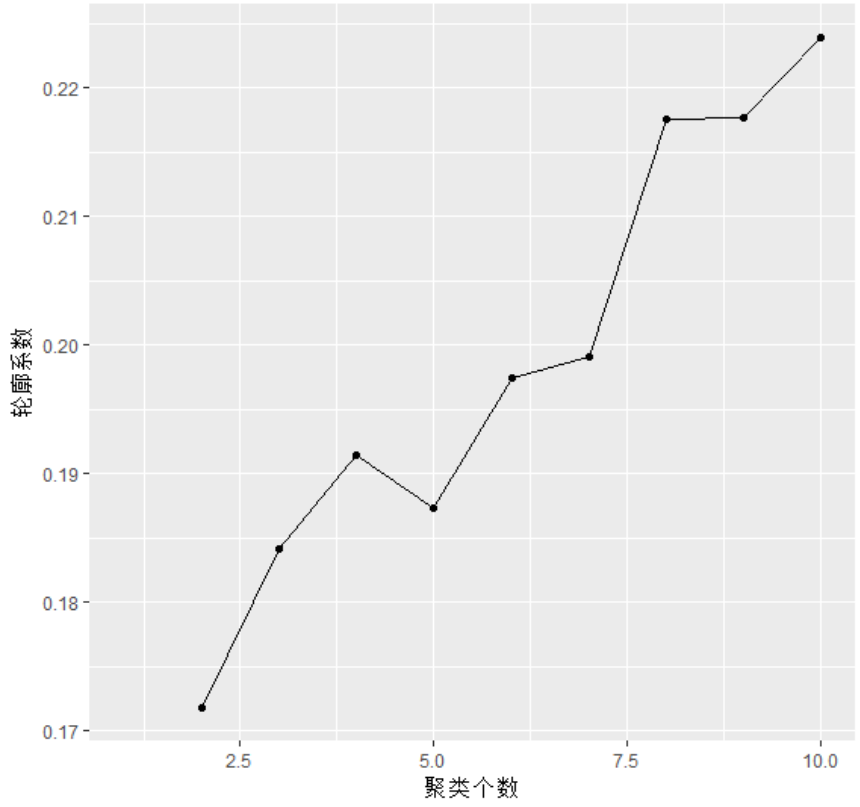

图 5 聚类个数与轮廓系数的散点图

图 6 为 10 类聚类中心点的平行坐标图，选取的坐标均为数值型变量。为了保证可视化效果，由于中心点样本在 PAY_4~PAY_6，BILL_AMT4~BILL_AMT6 这六个变量由于与前面同类别的变量的表现无较大差异，因此不予以显示。

可以看到，10 个中心点在 LIMIT_BAL（信用额度）变量上十分分散，PAY 类变量（过去 6 个月的付款状态）集中在-1 和 0（当月还清和循环信贷），两者有反比关系，说明信用额度大的人往往选择当月还清，额度小的人往往是选择循环信贷；

BILL_AMT 类变量(过去 6 个月账单数额)虽较为分散但也稳定，与 LIMIT_BAL 有正比关系，与 PAY 类变量有反比关系。

聚类中心样本在 PAY_AMT 类变量（过去 6 个月还款数额）上十分混乱，看不出明显关系。

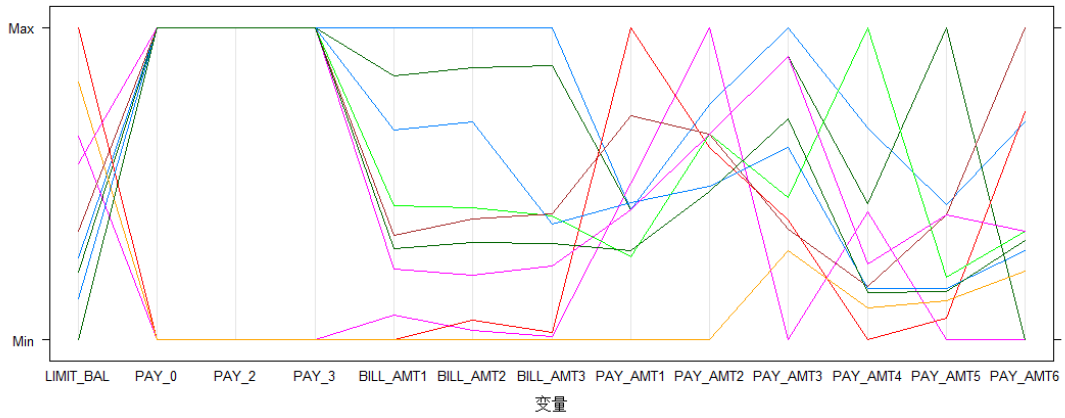

图 6 PAM 聚类中心的平行坐标图

t-SNE 是一种降维方法，它可以在保留聚类结构的前提下，将多维信息压缩到二维或三维空间中。借助 t-SNE 我们可以将 PAM 算法的聚类结果绘制出来，图 7 聚类图如下所示，可以很清楚地看到，聚类信用卡客户样本确实很分散，违约的客户分布在各个类别。

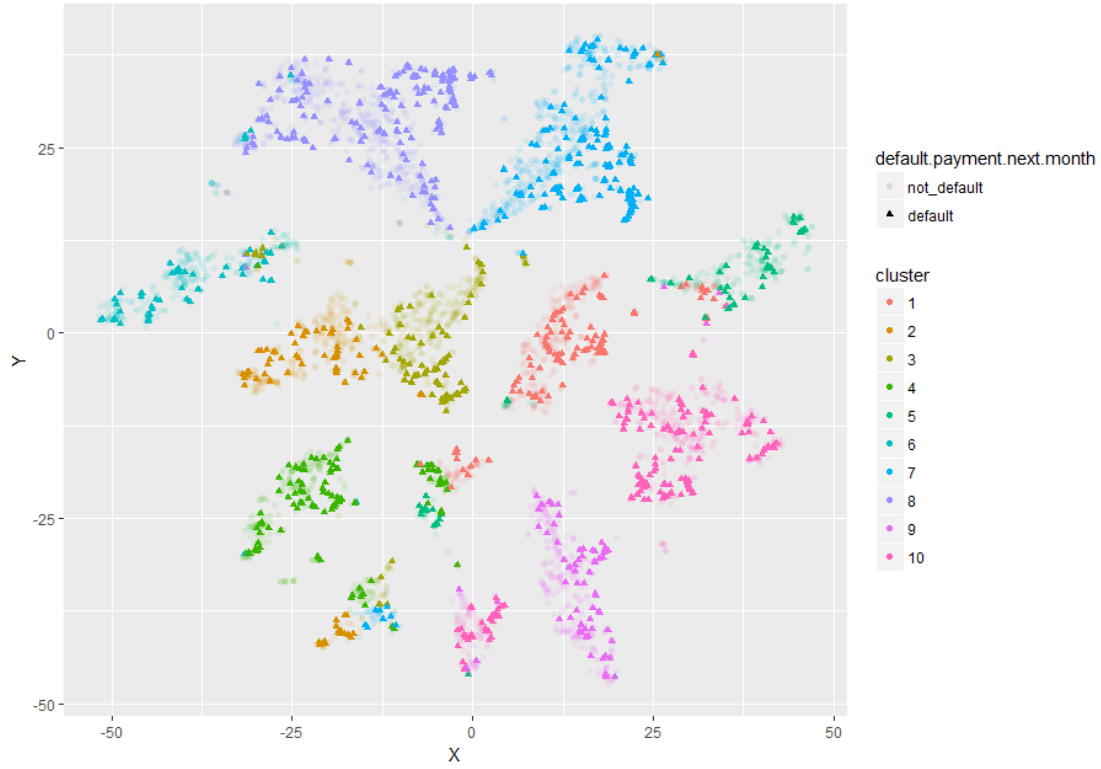

图 7 PAM 聚类结果可视化

2. K-prototype

K-prototype 是处理混合属性聚类的典型算法，继承 K-mean 算法和 K-mode 算法的思想。在 K-prototype 算法中混合属性的相异度分为数值型变量和分类变量分开求，然后相加。

K-prototype 的算法思想：两点间的距离被定义为 dist = d1 + w * d2，d1 是用 k-means 求得的连续型变量的距离，d2 是 k-modes 求得的分类变量间的差异，w 为权重。在聚类中使用的 w 为 0.5，但由于定量变量的相异度比定性变量的相异度大得多(数值型变量的标准差为 30991.82,分类变量的差异度为 0.54)，因此聚类结果几乎是由定量变量决定的。

图 8 聚类中心的平行坐标图如下所示，可见这与 PAM 聚类的结果不太一样，最明显的区别在于 K-prototype 的聚类中心在 PAY_AMT 类变量上变化较为分散和平缓，不像 PAM 聚类那样犬牙交错。

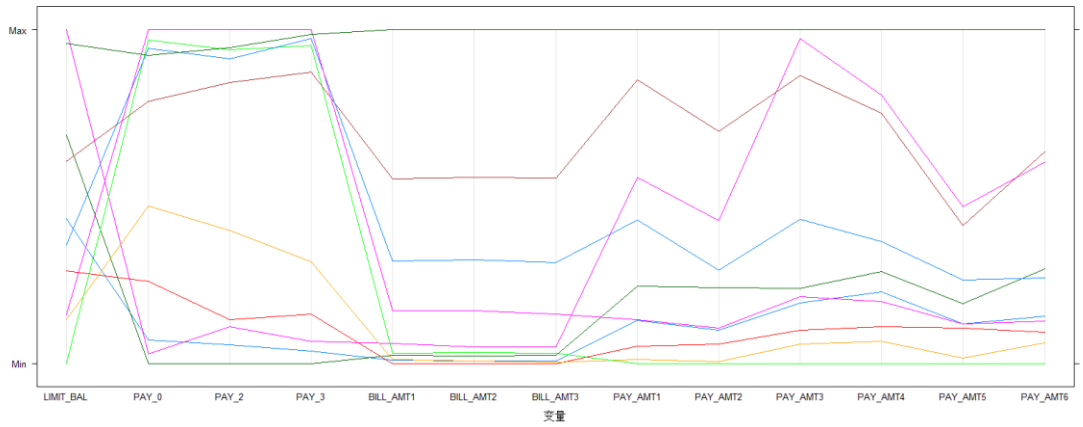

图 8 K-prototype 聚类中心的平行坐标图

3.K-means

既然聚类结果几乎是由定量变量决定的，猜想直接使用定量变量进行 k-means 聚类与 K-prototype 差别应当不大。下图为不同的聚类数时，kmeans 聚类结果的组内平方和（WCSS within-cluster sum of squares），组内平方和越小聚类的效果越好。图 9 也显示 kmeans 聚成 10 类比小于 10 类更好。

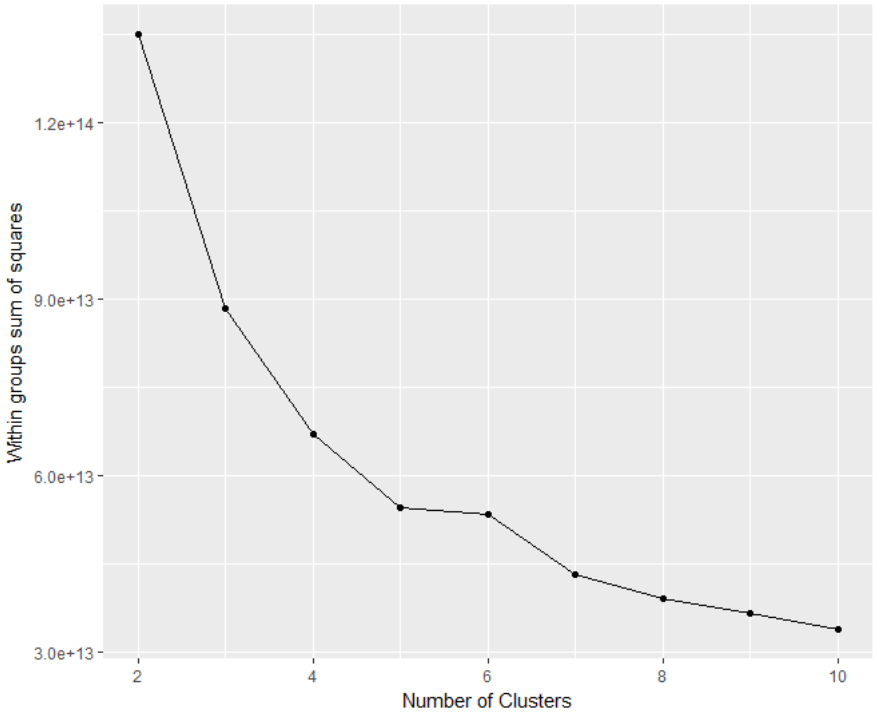

图 9 聚类个数与组内平方和的散点图

kmeans 聚类结果与 k-prototype 聚类的结果对比如下混淆矩阵图 10 所示，是一个稀疏矩阵，说明两个聚类结果重合度高，图 11kmeans 聚类中心的平行坐

标图与 k-prototype 聚类的差别不大，验证了聚类结果几乎是由定量变量决定的的想法。

```
> table(km_fit$cluster,kp_fit$cluster)
```

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	301	9	359	0
2	0	0	0	0	0	0	0	0	0	43
3	23	0	83	113	0	0	0	0	0	0
4	37	0	0	0	210	221	0	0	0	0
5	0	0	0	91	0	0	0	0	0	4
6	75	0	276	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	74	1125	0
8	27	562	0	0	223	0	0	0	0	0
9	0	133	0	0	0	0	6	539	0	0
10	15	10	16	0	0	0	425	0	0	0

图 10 kmeans 聚类结果与 k-prototype 聚类的结果混淆矩阵

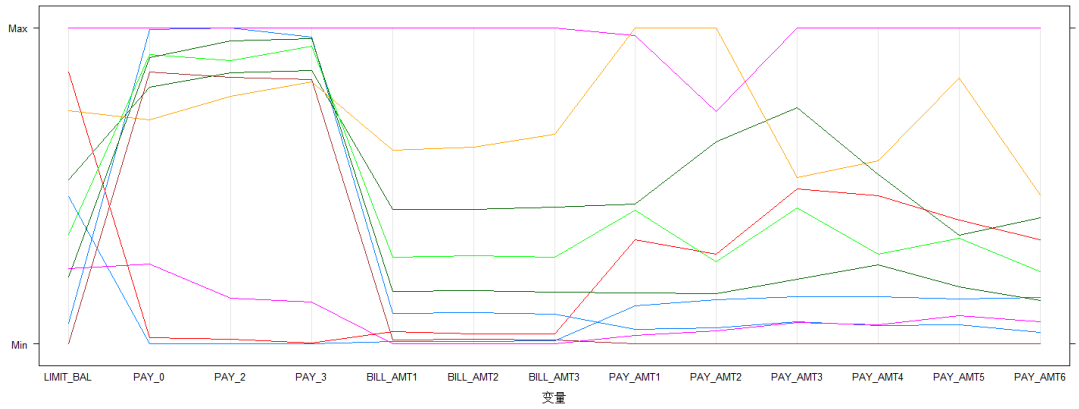

图 11kmeans 聚类中心的平行坐标图

4.聚类总结

由以上三种聚类方法可知，该数据中客户间差异较大，至少可以聚成 10 类，不同的类别的客户数据在数值型变量的取值差异大，聚类中心平行坐标图也反映出变量之间的关系。另外，混合型数据聚类中，对聚类结果起决定性作用的是标准差远远大于定性数据的差异度的定量变量，因此可以尝试只使用定量变量聚类，效果差别不会太大。

五、分类

随机抽取 60%的数据（18000 条样本）作为训练集，其他 40%的数据（12000 条样本）作为测试集。

1.Bagging

使用 500 棵树建立 bagging 模型，袋外错误率为 18.34%，测试误差为 18.26%，两者较为接近，说明模型并没有过拟合。图 12 中 ROC 曲线下的面积 AUC 为 0.666，说明尽管预测误差不大，但这是由于非违约客户占比太大导致，实际上识别违约的客户效果并不好，误分率高达 60.81%。

表 2bagging 的测试集混淆矩阵

cr_bag.pred	not_default	default	误分率
not_default	8754	554	0.0595
default	1637	1055	0.6081

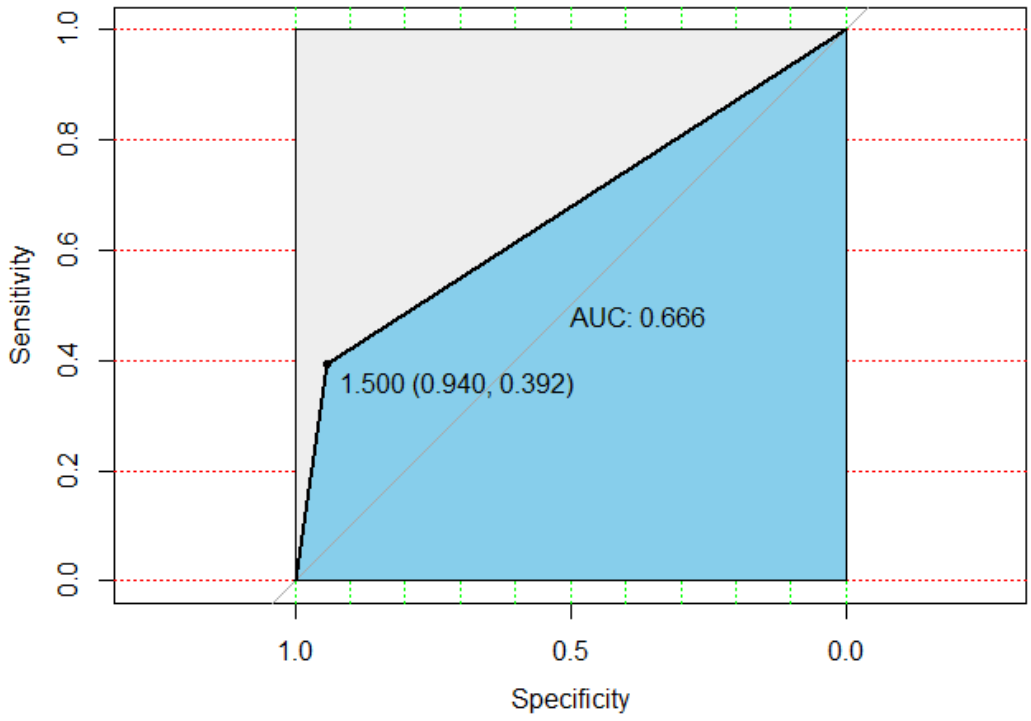

图 12 bagging 的测试集 ROC 曲线

2.随机森林

500 棵树，构造每棵树随机选取的特征数为$\sqrt{p}$ ，即 5 个，测试误差为 18.21%，比 bagging 的稍低，袋外错误率为 18.12%，图 13 中 ROC 曲线下的面积 AUC 为 0.663，比 bagging 稍高，混淆矩阵也显示识别的违约客户为 1026 人，误分率高达 61.89%，比 bagging 还高，说明随机森林识别违约客户没有 bagging 好。

表 3 随机森林的测试集混淆矩阵

cr_bag.pred	not_default	default	误分率
not_default	8789	519	0.0558
default	1666	1026	0.6189

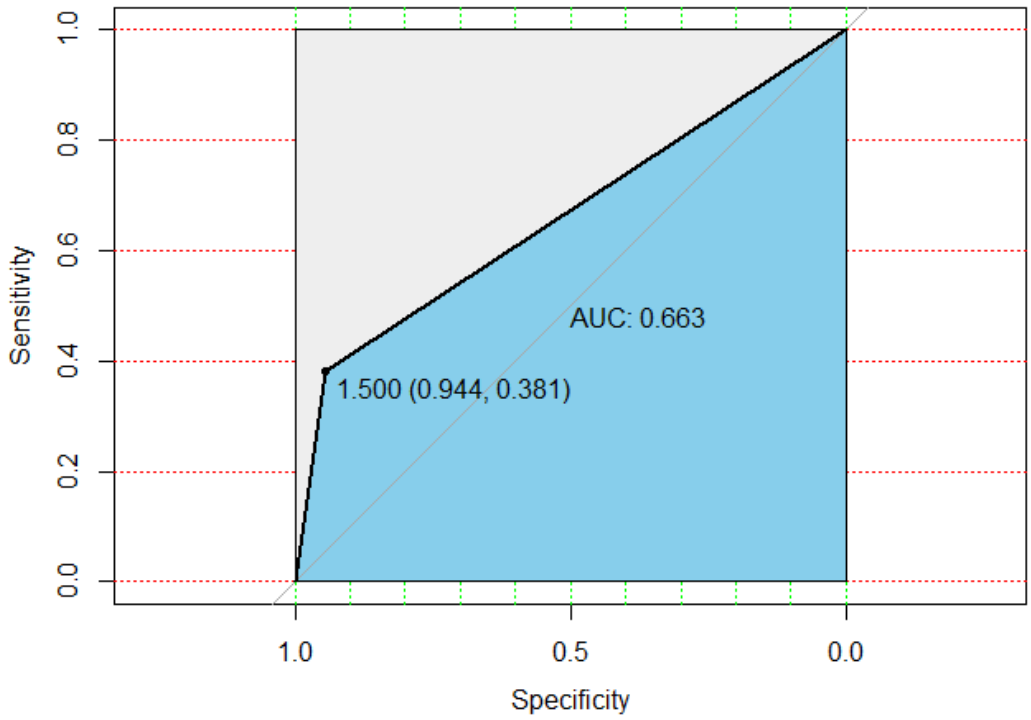

图 13 随机森林的测试集 ROC 曲线

由图 14 重要性图可以看出，PAY_0 变量最为重要，PAY_0 对降低均方误差和基尼系数的效果最好。

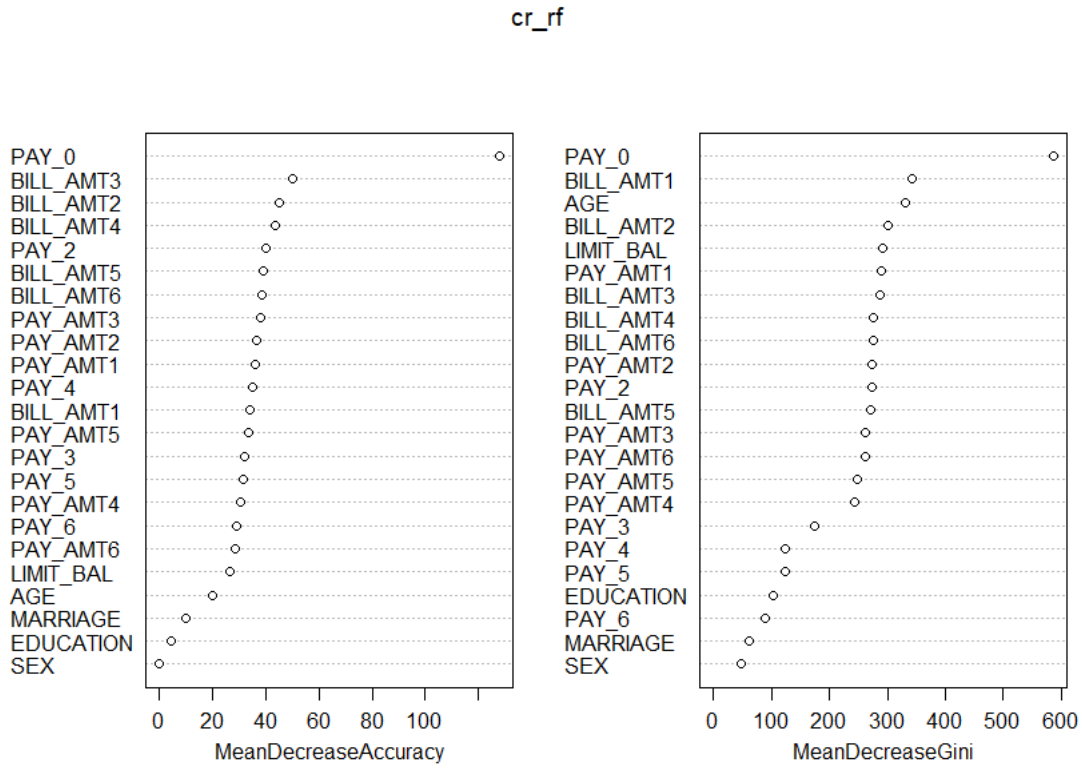

图 14 随机森林的变量重要性图

3.Adaboost

使用了 1000 棵树，树深为 4。由于 adaboost 的拟合值是连续型数字，因此需要找到阈值 p0 把训练拟合值划分为两类，好与真实值对比。

第一种 p0 是使预测精度最大化的阈值，求得当 p0=-0.68 时训练误差最大，为 17.85%,测试误差为 17.91%，比 bagging 和随机森林的结果都要好。此时图 15 显示 AUC 的值为 0.666，与 bagging 一样，算是较好的模型。但对违约客户识别率始终不高，错分率仍高达 61.44%。

表 4 第一种 p0 的 adaboost 模型的测试集混淆矩阵

cr_bag.pred	not_default	default	误分率
not_default	8813	495	0.0532
default	1654	1038	0.6144

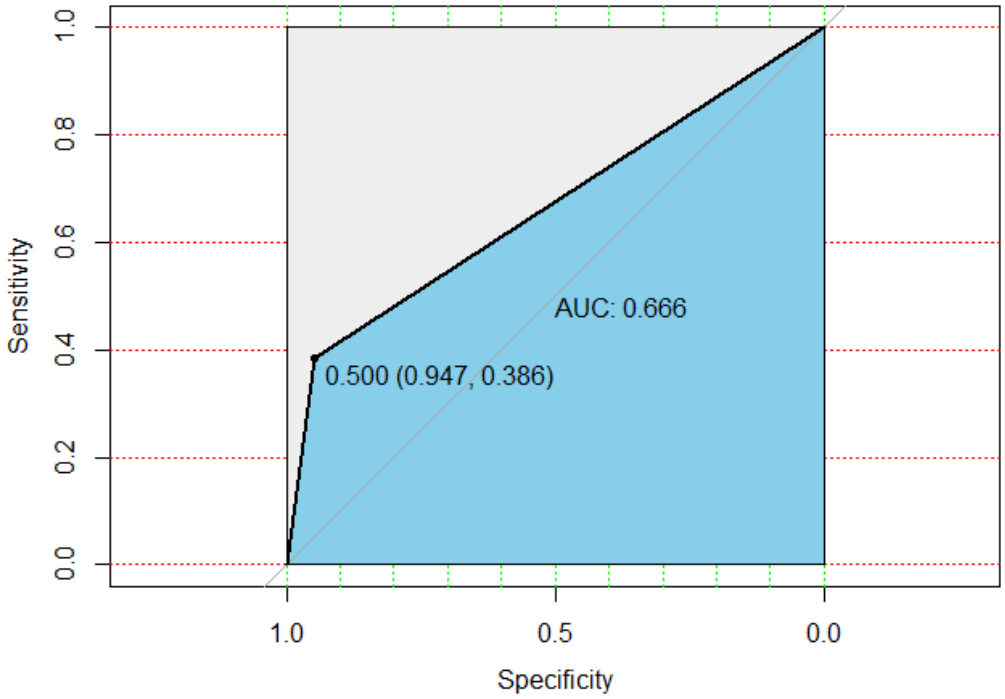

图 15 第一种 p0 的 adaboost 模型的测试集的 ROC 曲线

第二种 p0 是使 AUC 最大的阈值，图 16 左图中显示当 p0=-1.291 时训练集 AUC 值最大，此时测试误差为 23.05%，图 16 右图中设定 p0=-1.291 后测试集 AUC 高达 0.715，尽管测试误差较大，非违约客户错分率升高，但识别的违约客户变多，错分率大幅下降至 38.45%，说明这个模型对更关新违约客户的预测的银行等金融机构更有价值。

表 5 第二种 p0 的 adaboost 模型的测试集混淆矩阵

cr_bag.pred	not_default	default	误分率
not_default	7577	1731	0.1860
default	1035	1657	0.3845

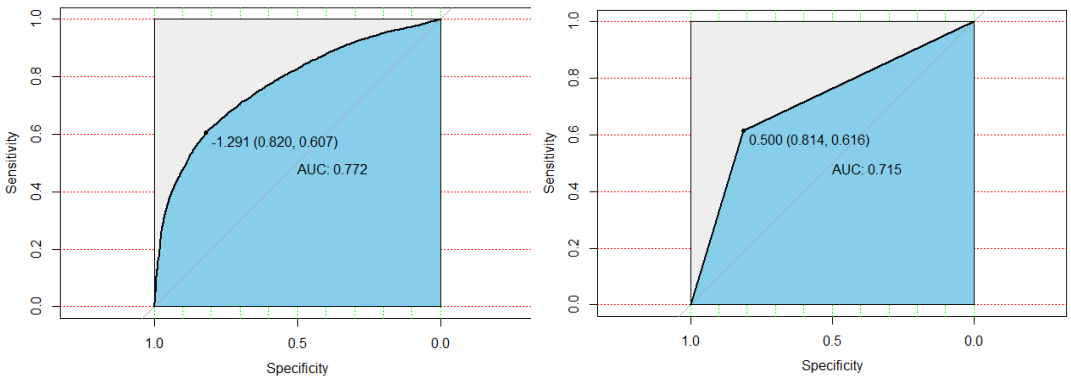

图 16 左图：训练集拟合值的 ROC 曲线；右图：第二种 p0 的 adaboost 模型的测试集 ROC 曲线

4.分类算法总结

由上述分类结果可知，如果盲目追求总体的预测准确率，会导致舍本逐末。因为对于银行等金融机构来说，他们更看重模型对违约客户的预测准确率，从这个意义上来说，adaboost 是相对于 bagging 和随机森林来说更好的模型。