

附录 1 应用问题示例

考虑在线贷款申请中劣质用户识别问题，基于已知的贷款申请人信用水平及所有申请人之间的关联信息，对新增贷款申请人的信用属性进行推断。这里将贷款申请人视为节点，假设有 n 个拥有信用记录的历史申请人以及 p 个没有信用记录的新增贷款申请人，形成规模为 $N = n + p$ 的网络。根据历史还款信息标记用户类别并作为响应变量， n 个历史申请人的响应变量是可以获取的，称该部分申请人为有标签用户；剩下 p 个新增申请人的响应变量是未知的，称为无标签用户。有标签用户中，如果节点 i 为劣质用户，即还款情况较差，则 $Y_i = 1$ ，否则 $Y_i = 0$ ；无标签用户中，响应变量为空值。个体间的网络连接定义如下，如果 i 和 j 相互有通话，则节点 i 和 j 相连；否则，则认为 i 和 j 不相连。图 1 是该数据的一个简单示例。图中，绿色正方形代表还款行为良好的优质用户，红色三角形代表有过长期逾期行为的劣质用户，两者都是有标签用户；白色圆形代表没有还款记录的无标签用户。图中 1 号节点周围的有标签用户均为劣质用户，推断 1 节点为劣质用户甚至可能是为劣质用户提供骗贷服务的“黑中介”。根据“近墨者黑”的思想，可以推断与 1 节点有直接关联的 2 节点和 3 节点很可能也是劣质用户。当然，实际生活中的情况会更加复杂，比如存在没有关联信息的孤立节点，或者响应变量为多分类变量。本文旨在通过网络中各个节点之间的连接信息和部分有标签用户的真实响应变量来预测剩余无标签用户的响应变量。

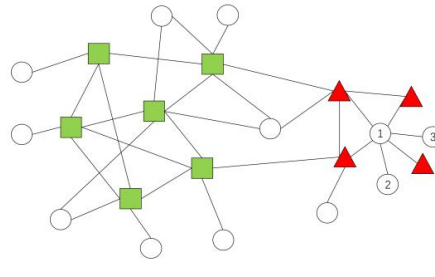


图 1 移动电话网络示例

附录 2 定理 2 的证明

在证明定理 2 之前，我们首先将伯恩斯坦不等式作为一个引理列出，其将在定理 2 的证明中使用。

引理 1. 对于独立随机变量 $X_1, X_2, \dots, X_n$ ，对任意 $i = 1, \dots, n$ ，若满足 $0 \leq |X_i| \leq \tau$ ，

$$S_n = \sum_{i=1}^n X_i,$$

$\mu_n = E(S_n)$ ， $V_n = Var(S_n)$ ，那么有

$$P(|S_n - \mu_n| \geq \varepsilon) \leq 2 \exp \left\{ -\frac{\varepsilon^2 / 2}{V_n + \tau \varepsilon / 3} \right\}.$$

对定理 2 的证明如下：

证明. 由文中的收敛性质可知， $\hat{Y}_U^{(a)} = (I - W_{UU})^{-1} W_{UL} Y_L$ 。根据 W 的定义以及泰勒展开公式，我们有 $(I - W_{UU})^{-1} = I + \sum (W_{UU})^m$ ，故 $\hat{Y}_U^{(a)}$ 可以展开为

$$\begin{aligned}\hat{Y}_U^{(a)} &= \left(I + \sum_{m=1}^{\infty} (W_{UU})^m \right) W_{UL} Y_L \\ &= W_{UL} Y_L + \sum_{m=1}^{\infty} (W_{UU})^m W_{UL} Y_L.\end{aligned}$$

我们接下来分两步来证明：首先证明随着 $N \rightarrow \infty$ ， $\sum (W_{UU})^m \rightarrow 0$ ；其次，再探讨 $W_{UL} Y_L$ 的收敛性质。

第一步，证明 $\sum (W_{UU})^m \rightarrow 0$ 。对于任意矩阵 $D = (D_{ij})$ ，定义 $\|D\|_{\max} = \max_{i,j} D_{ij}$ 。那么按照 W 的定义，我们有

$$\|W_{UU}\| \max_{i,j} \frac{a_{ij} + \alpha}{\sum_{j=1}^N a_{ij} + (N-1)\alpha} \leq \max_{i,j} \frac{a_{ij}}{\sum_{j=1}^N a_{ij}}.$$

我们首先证明存在常数 $c > 0$ ，使得

$$\min_i \sum_{j=1}^N a_{ij} \geq cN.$$

由于 $a_{ij} (i \neq j)$ 通过独立的伯努利分布生成，且它们之间相互独立只取值为 0 或 1，有 $E(a_{ij}) = p$ 或者 rp ；而矩阵 A 的对角线元素 $a_{ii} = 0$ 。则有

$$\sum_{j=1}^N E(a_{ij}) = \begin{cases} (n_1 - 1)p + n_0 rp, & \text{当 } Y_i = 1 \text{ 时} \\ (n_0 - 1)p + n_1 rp, & \text{当 } Y_i = 0 \text{ 时} \end{cases}$$

不失一般性，我们只考虑 $Y_i = 1$ 时的情况，那么 $\sum_{j=1}^N E(a_{ij}) = (n_1 - 1)p + n_0 rp$ 。根据引理 1

（伯恩斯

坦不等式）并取 $\tau = 1$ ，随着 $N \rightarrow \infty$ ，对任意 $\varepsilon > 0$ ，有

$$P(|\sum_{j=1}^N a_{ij} - (n_1 - 1)p - n_0 rp| \geq \varepsilon) \leq 2 \exp\left\{-\frac{\varepsilon^2 / 2}{V_n + \varepsilon / 3}\right\}.$$

其中， $V_n = \text{Var}\left(\sum_{j=1}^N a_{ij}\right)$ 。从而，

$$P\left(\left|\frac{1}{N} \sum_{j=1}^N a_{ij} - \frac{(n_1 - 1)p + n_0 rp}{N}\right| \geq \varepsilon\right) \leq 2 \exp\left\{-\frac{N^2 \varepsilon^2 / 2}{V_n + N\varepsilon / 3}\right\}.$$

(1)

由于 $V_n = \text{Var}\left(\sum_{j=1}^N a_{ij}\right) = O(N)$ ，故存在常数 $c_0 > 0$ ，使得式(1)可变为

$$P(|\frac{1}{N} \sum_{j=1}^N a_{ij} - \frac{(n_1-1)p + n_0rp}{N}| \geq \varepsilon) \leq 2 \exp\{-c_0 N \varepsilon\}.$$

(2)

由式(2)可得,

$$P(\frac{1}{N} \sum_{j=1}^N a_{ij} \leq \frac{(n_1-1)p + n_0rp}{N} - \varepsilon) \leq 2 \exp\{-c_0 N \varepsilon\}.$$

(3)

式(3)对任意 ε 都成立。定义 $c = \frac{(n_1-1)p + n_0rp}{2N}$, 当 $\varepsilon = c$ 时, 运用 Bonferroni 不等式, 我

们有

即以概率 1 成立

$$\min_i \sum_{j=1}^N a_{ij} \geq cN.$$

故有

$$\begin{aligned} \|W_{uu}\|_{\max} &\leq \max_i \frac{1}{\sum_{j=1}^N a_{ij}} \leq \frac{1}{cN} \\ \| (W_{uu})^2 \|_{\max} &= \|W_{uu} \cdot W_{uu}\|_{\max} \leq \frac{1}{cN} \frac{n_u}{cN} \\ &\dots \\ \| (W_{uu})^m \|_{\max} &\leq \frac{1}{cN} \left(\frac{n_u}{cN} \right)^{m-1} \end{aligned}$$

又因为 $n_U = O(N^\gamma)$, $0 < \gamma < 1$, 故 $\frac{n_U}{cN} < 1$, 则有

$$\begin{aligned} \left\| \sum_{m=1}^{\infty} (W_{UU})^m \right\| &\leq \frac{1}{cN} \sum_{m=1}^{\infty} \left(\frac{n_U}{cN} \right)^{m-1} = \frac{1}{cN \left(1 - \frac{n_U}{cN} \right)} \\ &= \frac{1}{cN - n_U}. \end{aligned}$$

故随着 $N \rightarrow \infty$, $\sum_{m=1}^{\infty} (W_{UU})^m \rightarrow 0$.

第二步, 我们讨论 $w_{UL} Y_L$ 的理论性质。由第一步可知, $\hat{Y}_U^{(\alpha)} \rightarrow w_{UL} Y_L$ 。从而对任意 $i \in S_1$, 有

$$\hat{Y}_i^{(\alpha)} \rightarrow \sum_{j \in S_0} w_{ij} Y_j, w_{ij} = \frac{a_{ij} + \alpha}{\sum_{j=1}^N a_{ij} + (N-1)\alpha}.$$

对任意 $i \in S_1$ ，我们按其真实标签的取值分为两种情况： $Y_i = 1$ 和 $Y_i = 0$ 。

若 $Y_i = 1$ ，那么根据第一步的式(2)，有

$$\frac{1}{N} \sum_{j=1}^N a_{ij} \rightarrow \frac{(n_1 - 1)p + n_{0rp}}{N}.$$

从而，

$$\hat{Y}_i^{(\alpha)} \rightarrow \frac{1}{(n_1 - 1)p + n_{0rp} + (N - 1)\alpha} \sum_{j \in S_0} (a_{ij} + \alpha) Y_j.$$

设有标签节点 S_0 里面标签为 1 的节点数为 n_{L_1} ，标签为 0 的节点数为 n_{L_0} ，则 $n_L = n_{L_1} + n_{L_0}$ ，

故

$$\hat{Y}_i^{(\alpha)} \rightarrow \frac{1}{(n_1 - 1)p + n_{0rp} + (N - 1)\alpha} \sum_{j \in S_0} (a_{ij} + \alpha) Y_j.$$

由于 Y_j 仅取值为 0 或 1，故与第一步中的证明类似，

$$\sum_{j \in S_0} E(a_{ij} Y_j) = \sum_{j \in S_0, Y_j=1} p Y_j + \sum_{j \in S_0, Y_j=0} r p Y_j = n_{L_1} p.$$

进一步可得

$$\sum_{j \in S_0} a_{ij} Y_j \rightarrow n_{L_1} p,$$

$$\hat{Y}_i^{(\alpha)} \rightarrow \frac{n_{L_1} (p + \alpha)}{(n_1 - 1)p + n_{0rp} + (N - 1)\alpha}.$$

若 $Y_i = 0$ ，同理可得

$$\hat{Y}_i^{(\alpha)} \rightarrow \frac{n_{L_1} (rp + \alpha)}{(n_0 - 1)p + n_{1rp} + (N - 1)\alpha}.$$

令 $c_1 = \frac{n_{L_1} (p + \alpha)}{(n_1 - 1)p + n_{0rp} + (N - 1)\alpha}$ ， $c_2 = \frac{n_{L_1} (rp + \alpha)}{(n_0 - 1)p + n_{1rp} + (N - 1)\alpha}$ ，又由于 $N = n_1 + n_0$ ，

$0 < r < 1$ ，则通过计算化简可得

$$\frac{c_1}{c_2} = \frac{(n_0 - 1)p^2 + n_0 \alpha p + (n_0 - 1)\alpha p}{(n_0 - 1)r^2 p^2 + n_0 \alpha r p + (n_0 - 1)\alpha r p} > 1,$$

即 $c_1 > c_2 > 0$ 。

综上，对任意无标签节点 $i \in S_1$ ，随着 $N \rightarrow \infty$ ，若 $Y_i = 1$ ，则 $\hat{Y}_i^{(\alpha)} \rightarrow c_1$ ；若 $Y_i = 0$ ，则 $\hat{Y}_i^{(\alpha)} \rightarrow c_2$ 。故对任意 $\alpha \in (0, 1)$ ，一定存在 $\tau \in (c_2, c_1)$ ，比如 $\tau = (c_1 + c_2)/2$ ，使得 $P(\hat{Y}_U^{(\alpha, \tau)} = Y_U) \rightarrow 1$ ，其中 $\hat{Y}_U^{(\alpha, \tau)} = I(\hat{Y}_U^{(\alpha)} > \tau)$ 。这就证明了本文的结论。

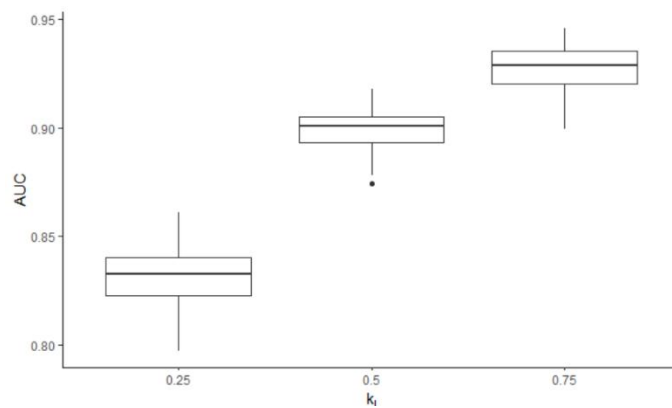
附录 3 多分类问题数值模拟

正文第三部分提出本文算法也适用于多分类问题，为此，本文考虑了三分类问题的数值模拟实验。首先生成 N 个节点：将网络中 N 个节点归为 3 个子网络，响应变量取值分别为 1, 0, -1，样本量分别为 $N/2$ ， $N/4$ ，和 $N/4$ 。接下来生成邻接矩阵 A ：如果节点 i 和 j 在同一个子网络中，则由概率为 p 的伯努利分布独立生成连接边（边取值为 1 代表两个节点之间有联系，边取值为 0 代表两个节点之间无联系）；如果节点 i 和 j 在不同的子网络中，则由概率为 rp 的伯努利分布独立生成连接边。在本例中，和二分类问题的设定类似，假设 $r = 0$ ， $p = 0.1$ 。因此，同一子网络内的节点间有连接边的概率大于不同子网络之间有连接边的概率。从整个网络中随机选择 $0.4N$ 个节点设置为无标签节点，从而得到有标签节点响应变量 Y_l 和无标签节点响应变量 Y_u 。针对度修正随机分块模型，其度修正参数向量 $\omega = (\omega_1, \dots, \omega_N)$ 的生成方式和二分类问题相同。同样采用 ACC、AUC 和 F 统计量作为多分类预测效果的衡量指标。其中，AUC 值为将该三分类问题转变为三个二分类问题后所计算的 AUC 值的平均值。模拟结果如表 1 所示。表 1 的结果和正文表 1~4 中的结果具有同样的趋势，这表明本文方法同样适用于多分类问题。

表 1 不同样本量 N 的三分类问题模拟结果汇总

N	随机分块模型			度修正随机分块模型		
	AUC	ACC	F	AUC	ACC	F
1000	0.872	0.843	0.862	0.826	0.803	0.821
2000	0.923	0.885	0.908	0.877	0.867	0.870
3000	0.944	0.921	0.927	0.905	0.884	0.889

附图 1 AUC 箱线图



附图 1 实际数据在不同 m 下预测 AUC 的分布情况。横轴展示 $1-m$ 的取值。