基于多源数据融合的个人信用评分研究*

方匡南 赵梦峦

内容提要: 随着信息技术的发展,数据来源越来越多,虽然可以更加精准、科学地刻画个人信用状况,但由于数据来源多、结构复杂等问题,给传统的征信技术带来了挑战。本文提出了基于多源数据融合的个人信用模型,可以同时对多个数据集进行建模和变量选择,同时考虑了数据集间的相似性和异质性。通过模拟实验发现,本文所提出的整合模型在变量选择和分类效果方面都具有明显的优势。此外,将整合模型应用于城市和农村两个数据集的个人信用评分中发现,整合模型在实际应用中也有很好的表现。

关键词: 多源数据; 整合分析; Logistic 回归; 信用评分
DOI: 10.19343/j.cnki.11-1302/c.2018.12.008
中图分类号: C812 文献标识码: A 文章编号: 1002-4565(2018)12-0092-10

A Study on Credit Scoring Based on Multi-source Data Integration

Fang Kuangnan & Zhao Mengluan

Abstract: With the development of internet technology, data sources become diversified. It is possible to get more accurate personal credit status on one hand, but on the other hand, due to multi data sources and complicated data structure, it is a great challenge to the traditional credit collection techniques. This paper proposes a new credit scoring model based on multi-source data integration. It can simultaneously build up models and select variables using multiple data sets, taking stock of the homogeneity and heterogeneity of the data sets, but also considering the similarity between the data sets. It is found in the simulation that, the integrated model proposed has a significant advantage in both variable selection and effective classification. Finally, the urban and rural data sets in China are applied to the integrated personal credit scoring model.

Key words: Multi-source Data; Integrative Analysis; Logistic Regression; Credit Scoring

一、引言

近年来,随着互联网金融的发展,P2P 网贷、车贷、消费贷、现金贷等网络信用贷款发展迅速,借助互联网的优势,可以足不出户地完成贷款申请。但网络交易的虚拟性,也容易产生欺诈和欠款不还等信用风险。我国商业银行和互联网金融公司的风险管理技术面临着众多挑战,尤其是针对多

* 本文获国家自然科学基金面上项目"广义线性模型的组变量选择及其在信用评分中的应用"(71471152)、全国统计科学研究重点项目"大数据下的信用评分研究"(2015629)以及中央高校基本科研业务费专项资金"多源异构大数据的整合分析研究"(20720171095)的资助。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

源海量数据的征信技术还比较欠缺,给我国的金融安全带来一定的隐患。

传统的信用决策系统主要依赖于经验丰富的信贷专家的主观判断,评判标准难以做到科学、客观和统一。为了降低信贷决策中的主观因素,基于统计模型的信用评估方法受到越来越多的关注。常用的方法有判别分析、神经网络、支持向量机和 Logistic 回归模型等,其中最具代表性的是 Logistic 回归模型,由于其预测准确率高、计算简单、解释能力强等特点被学术界广泛采用(胡心瀚等,2012[1]; 方匡南等,2014[2]) 。Wiginton(1980) [3] 将 Logistic 模型与判别分析对比,认为 Logistic 模型比判别分析效果更好。West(2000) [4] 也认为神经网络模型并不比 Logistic 预测效果好。李志辉和李萌(2005) [5] 在研究我国商业银行信用风险的识别时,比较分析了 Fisher 线性判别分析、BP 神经网络和 Logistic 方法的效果,研究表明 Logistic 模型具有更强的信用风险识别和预测能力。

Logistic 模型虽然简单易操作,但也存在局限性,一方面如果模型中包含过多变量,变量之间往往存在多重共线性,导致模型解释性和预测精度降低(方匡南等,2014) ; 另一方面,模型中选入过多无关变量,不仅对理解变量间的关系产生干扰,还会因为收集无关变量信息而浪费大量的人力物力(王大荣和张忠占,2012) [6] 。因此,很多学者将变量选择方法应用于我国的个人信用评分中。方匡南等(2014) 将 Lasso 惩罚应用到 Logistic 模型,构建了基于 Lasso-logistic 的个人信用评分模型,同时进行变量选择和建模,不仅提高了模型的解释性也提升了模型的预测精度。王小燕等(2014) [7] 则提出基于 aSGL-Logistic 回归,对分组结构变量进行组内和组间双层选择,并应用于我国个人信用评分中。胡小宁等(2015) [8] 研究了基于 Group MCP Logistic 模型的个人信用评价。方匡南等(2016) [9] 将弹性网结构引入 Logistic 回归,提出带网络结构的 Logistic 模型并应用到我国的信用风险甄别中。

随着互联网技术的发展,数据来源越来越多,评估信用的变量维度越来越高,可以更加精准、科学地刻画信用状况,但由于数据来源多、数据结构复杂、存在缺失值、数据非平衡等问题,对传统的征信技术带来了极大的挑战。再比如来自城市和农村的征信数据,由于我国城市和农村的经济发展程度、收入、消费等各方面均存在明显差异,信用消费和信用卡的普及程度不同,消费观念和消费能力存在差异等,这些差异可能会对信用违约产生不同的影响。对于这类多源数据,如果按照传统的处理方法,将所有数据集简单合并成一个数据集,则数据间的异质性会被忽略; 但如果针对不同的数据集分别独立建模,则又可能会忽略数据集间的关联性(马双鸽等,2015) [10] 。只有同时考虑数据集间的异质性和关联性,才能更有效地利用所有数据,获得更可靠的结果。因此需要对多源数据进行整合分析,该方法起源于 20 世纪 60 年代, 可以把不同来源、不同格式、不同特点的数据进行融合建模。相对于单一数据集模型,该方法整合了更多的原始信息,能解决因不同来源数据的差异而引起的建模不稳定,在模型解释性和预测方面都具有显著优势。多源数据的整合分析,可以分为两类,一类是“变量”方向的整合,即不同数据集样本相同而变量不同,整合不同来源的变量信息,比如信用评分时整合了贷款记录数据、网购数据、公安法院数据等; 另一类是“样本”方向的整合,即样本是不同的但变量是一样的,整合不同来源的样本信息,比如城市和农村居民的征信数据。本文所研究的主要是“样本”方向的整合建模。

本文的创新主要有: ①提出了基于多源数据融合的 Logistic 回归模型,通过“损失函数 + 变量选择惩罚函数 + 相似性惩罚函数”方式综合不同来源的数据集,从统计角度考虑数据集的异质性和关联性,并同时实现高维数据的系数估计和变量选择; ②提出的整合模型一方面可以选出每个数据集上的显著变量,另一方面又考虑不同数据集间的关联性,比如同一变量在不同数据集中具有系数符号相似性等特点; ③针对目标函数不可导等问题,给出了可行的优化算法; ④将本文所提出的整合模型应用到我国的个人信用评分中,可以对多源数据集同时建模分析。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

二、模型与算法

(一) 模型

假设有 M 个数据集，每个数据集有 d 个解释变量，第 m 个数据集的样本量为 n_m，被解释变量 y^m 为二元变量，解释变量 X^m 为 $n_m \times d$ 矩阵。本文基于 Logistic 回归，即：

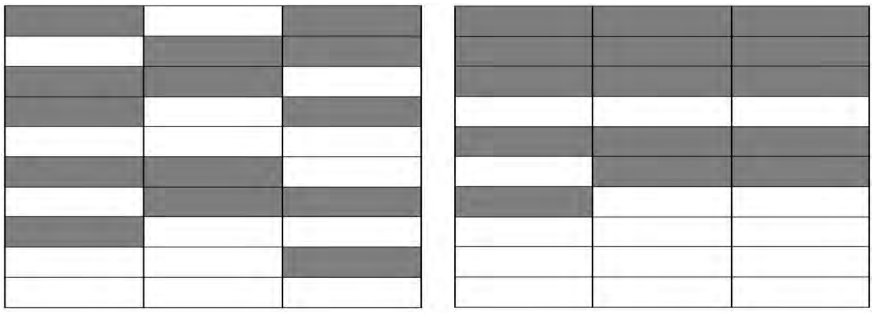
$$\log\left(\frac{p_i^m}{1-p_i^m}\right) = \beta_0^m + \beta_1^m X_{1i}^m + \cdots + \beta_d^m X_{di}^m, m = 1, \cdots, M$$

其中 $p_i^m = \Pr(y_i^m = 1) = 1/[1 + \exp(-(\beta_0^m + \sum_{j=1}^{d} x_{ij}^m \beta_j^m))]$，$\beta_0^m$ 为第 m 个模型的截距项，$\beta_j^m (j = 1, \cdots, M)$ 为第 m 个模型第 j 个变量的系数。对 M 个数据集同时估计系数和筛选变量，采用“损失函数 + 惩罚函数”形式。提出了式(1)的目标函数：

$$L(\beta; \lambda_1, \lambda_2, a, b) = \sum_{m=1}^{M}\left\{-\frac{1}{n_m}\left\{\sum_{i=1}^{n_m}[y_i^m(\beta_0^m + \sum_{j=1}^{d} x_{ij}^m \beta_j^m) - \log(1 + \exp(\beta_0^m + \sum_{j=1}^{d} x_{ij}^m \beta_j^m))]\right\}\right\} + \sum_{j=1}^{d} p\left(\sum_{m=1}^{M} p(|\beta_j^m|; \lambda_1, a); \lambda_1, b\right) + \frac{\lambda_2}{2}\sum_{j=1}^{d}\sum_{1 \leqslant l, m \leqslant M}(\operatorname{sgn}(\beta_j^m) - \operatorname{sgn}(\beta_j^l))^2 \quad (1)$$

式(1)的第一项是损失函数，即 Logistic 对数似然函数在 M 个数据集加权求和的相反数，之所以对每个数据集按样本量倒数进行加权求和是为了消除不同数据集因样本量不同产生的影响，之所以在 Logistic 对数似然函数前加负号是因为要求使得式(1)最小化的解；第二项是复合 MCP 惩罚函数，即组内惩罚和组间惩罚均采用 MCP 惩罚函数，可以用来双层变量选择，组间惩罚函数选择在所有数据集上均显著的变量，组内惩罚函数选择这些变量在哪些数据集上显著，需要说明的是除了本文使用的是复合 MCP 惩罚函数，也可以使用其他惩罚函数；第三项是符号相似惩罚函数，该惩罚函数的作用是对变量系数的符号差异进行惩罚，鼓励同一变量在不同数据集上的系数符号相似。其中，$\operatorname{sgn}(t)$ 是符号函数，在 $t > 0$、$t = 0$、$t < 0$ 的情况下，分别有 $\operatorname{sgn}(t) = 1$、0、$-1$。

如图 1 所示，假设有 3 个数据集，每列对应一个数据集，每行对应一个变量，假设前 10 个变量均是显著的，其中阴影表示变量的系数为正，白色表示变量的系数为负。图 1(a) 是加符号相似性惩罚前，不同的数据集上变量系数的符号差异较大，图 1(b) 是加符号相似性惩罚后，不同的数据集上的系数符号趋于相似。



(a) 加符号相似性惩罚前　　(b) 加符号相似性惩罚后

图 1　符号相似性整合分析示意图

由于 Logistic 回归的损失函数没有显式解，所以本文采用 MM 算法进行二次近似。由于总损失函数为 M 个数据集损失函数的加权求和，本文以第 m 个数据集的损失函数 $L(X^m, y^m; \beta^m)$ 为例，它

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

关于 β^m 的一阶导数和二阶导数分别为:

$$\frac{\partial L(X^m, y^m; \beta^m)}{\partial \beta^m} = -\sum_{i=1}^{n} y_i^m x_i^m + \sum_{i=1}^{n} p_i^m x_i^m = X^{m^T}(p^m - y^m) \tag{2}$$

$$\frac{\partial^2 L(X^m, y^m; \beta^m)}{\partial \beta^m \partial \beta^{m^T}} = X^{m^T} W X^m \tag{3}$$

其中 W 为对角矩阵,对角元素为 $p_i^m(1 - p_i^m)$。令 $\eta^m = \beta_0^m + X^m\beta^m$,同时令 $v = \max_i \sup_{\eta^m}\{\nabla^2 L_i(\eta^m)\}$,因为 $\nabla^2 L_i(\eta^m) = p_i^m(1 - p_i^m)$,所以矩阵 W 的对角元素为 v,而且可得 $v = 1/4$。

根据 MM 算法,本文可以对损失函数部分进行迭代,直至收敛,将损失函数在当前的估计值 $\tilde{\beta}$ 处进行二次泰勒展开(Breheny,2015)[11],可得:

$$\tilde{L}(\beta) = (\beta - \tilde{\beta})^T X^T(p - y) + (\beta - \tilde{\beta})^T \frac{\partial L}{\partial \beta \partial \beta^T}(\beta - \tilde{\beta}) \propto \frac{v}{2}(X\tilde{\beta} + \frac{y - p}{v} - X\beta)^T(X\tilde{\beta} + \frac{y - p}{v} - X\beta) \tag{4}$$

对式(4)进行整理替换,可将对数似然函数的负向函数替换为式(5)的二次损失函数,并对其进行优化:

$$R^m(\beta_0^m, \beta^m \mid \tilde{\beta}_0^m, \tilde{\beta}^m) = \frac{\nu}{2}\sum_{i=1}^{n_m}(z_i^m - \beta_0^m - \sum_{j=1}^{d} x_{ij}^m\beta_j^m)^2 + C(\tilde{\beta}_0^m, \tilde{\beta}^m) \tag{5}$$

其中,$z_i^m = \tilde{\beta}_0^m + \sum_{j=1}^{d} x_{ij}^m\tilde{\beta}_j^m + \frac{1}{\nu}[y_i^m - \tilde{p}(x_i^m)]$ 和 $\tilde{p}(x_i^m) = 1/[1 + \exp(-(\tilde{\beta}_0^m + \sum_{j=1}^{d} x_{ij}^m\tilde{\beta}_j^m))]$ 是根据迭代中已获得的估计值 $(\tilde{\beta}_0^m, \tilde{\beta}^m)$ 计算所得,$C(\tilde{\beta}_0^m, \tilde{\beta}^m)$ 是只与当前估计值 $(\tilde{\beta}_0^m, \tilde{\beta}^m)$ 有关、而与 (β_0^m, β^m) 无关的数,在迭代时可不予考虑。

对于符号惩罚函数部分,由于符号函数 $\mathrm{sgn}(t)$ 不连续,无法求导,难以对其进行最优化求解,为解决该问题,借鉴 Fang 等(2018)[12] 的研究,将其近似为式(6)的惩罚函数:

$$\frac{\lambda_2}{2}\sum_{j=1}^{d}\sum_{1\leqslant l,m\leqslant M}\left(\frac{\beta_j^m}{\sqrt{\beta_j^{m2} + \tau^2}} - \frac{\beta_j^l}{\sqrt{\beta_j^{l2} + \tau^2}}\right)^2 \tag{6}$$

其中 $\tau > 0$ 是一个很小的常数。

通过对损失函数和惩罚函数的二次近似,对式(1)的最优化问题就变成了对式(7)的最优化问题:

$$L(\beta) \approx \sum_{m=1}^{M}\frac{1}{n_m}\left\{\frac{1}{8}\sum_{i=1}^{n_m}(z_i^m - \beta_0^m - \sum_{j=1}^{d} x_{ij}^m\beta_j^m)^2 + C(\tilde{\beta}_0^m, \tilde{\beta}^m)\right\} + \sum_{j=1}^{d} p(\sum_{m=1}^{M} p(|\beta_j^m|; \lambda_1, a); \lambda_1, b) + \frac{\lambda_2}{2}\sum_{j=1}^{d}\sum_{l\neq m}\left(\frac{\beta_j^m}{\sqrt{\tilde{\beta}_j^{m2} + \tau^2}} - \frac{\tilde{\beta}_j^l}{\sqrt{\tilde{\beta}_j^{l2} + \tau^2}}\right)^2 \tag{7}$$

(二)模型计算

本文采用 Yuan 和 Lin(2006)[13] 提出的组坐标下降法(Group Coordinate Descent, GCD)来进行优化求解,GCD 是坐标下降法(Coordinate Descent, CD)在组结构下的拓展,迭代过程中,先固定其他 $d-1$ 组参数,对 β_j 在 M 个数据集上估计,依次对 d 个变量进行优化,通过迭代,每次只优化一组参数,直到所有参数都收敛到给定精度为止。本文基于 GCD 对目标函数进行求解,具体求解过程如下所述:

本文以第 m 组数据集的第 j 个变量的优化为例,固定其他变量参数不变,也可以忽略加在其他参数上的惩罚,此时本文的目标函数为:

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$L_j^m = \frac{1}{n_m}\left\{\frac{1}{8}\sum_{i=1}^{n_m}\left(z_i^m - \beta_0^m - \sum_{j=1}^{d} x_{ij}^m\beta_j^m\right)^2 + C\left(\tilde{\beta}_0^m, \tilde{\beta}^m\right)\right\} + p\left(\sum_{m=1}^{M} p\left(|\beta_j^m|;\lambda_1, a\right);\lambda_1, b\right) + \frac{\lambda_2}{2}\sum_{l\neq m}\left(\frac{\beta_j^m}{\sqrt{\tilde{\beta}_j^{m2} + \tau^2}} - \frac{\tilde{\beta}_j^l}{\sqrt{\tilde{\beta}_j^{l2} + \tau^2}}\right)^2 \quad (8)$$

首先求解截距项，令 L_j^m 对 β_0^m 求导等于 0，即:

$$-\frac{1}{4n_m}\sum_{i=1}^{n_m}\left(z_i^m - \beta_0^m - \sum_{j=1}^{d} x_{ij}^m\beta_j^m\right) = 0 \quad (9)$$

可得:

$$\hat{\beta}_0^m = \frac{1}{n_m}\sum_{i=1}^{n_m}\left(z_i^m - \sum_{j=1}^{d} x_{ij}^m\tilde{\beta}_j^m\right) \quad (10)$$

接下来令 L_j^m 对 β_j^m 求一阶导等于 0，由于目标函数式(8)的三个部分相对独立，可分别先求这三部分的导数然后相加，接下来 MCP 惩罚部分对 β_j^m 求偏导，导数为:

$$\tilde{\lambda}_{jm}\partial|\beta_j^m|/\partial\beta_j^m \quad (11)$$

其中，$\tilde{\lambda}_{jm} = p'\left(\sum_{m=1}^{M} p\left(|\tilde{\beta}_j^m|;\lambda_1, a\right);\lambda_1, b\right)p'\left(|\tilde{\beta}_j^m|;\lambda_1, a\right)$。

对符号惩罚函数部分用完全平方公式展开如式(12):

$$\frac{\lambda_2}{2}\sum_{m\neq l}\left(\frac{\beta_j^m}{\tilde{\beta}_j^{m2} + \tau^2} - \frac{2\beta_j^m\tilde{\beta}_j^l}{\sqrt{\tilde{\beta}_j^{m2} + \tau^2}\sqrt{\tilde{\beta}_j^{l2} + \tau^2}} + \frac{\tilde{\beta}_j^l}{\tilde{\beta}_j^{l2} + \tau^2}\right) \quad (12)$$

式(12)对 β_j^m 求偏导为:

$$\frac{\lambda_2}{2}\sum_{m\neq l}\left(\frac{2}{\tilde{\beta}_j^{m2} + \tau^2}\beta_j^m - \frac{2\tilde{\beta}_j^l}{\sqrt{\tilde{\beta}_j^{m2} + \tau^2}\sqrt{\tilde{\beta}_j^{l2} + \tau^2}}\right) = \frac{\lambda_2(M-1)}{\tilde{\beta}_j^{m2} + \tau^2}\beta_j^m - \frac{\lambda_2}{\sqrt{\tilde{\beta}_j^{m2} + \tau^2}}\sum_{m\neq l}\frac{\tilde{\beta}_j^l}{\sqrt{\tilde{\beta}_j^{l2} + \tau^2}} \triangleq A\beta_j^m + B \quad (13)$$

将式(9)、式(11)、式(13)的导数加总在一起，则目标函数 L_j^m 对 β_j^m 的导数如式(14):

$$-\frac{1}{4n_m}x_j^{mT}\left(z^m - \tilde{\beta}_0^m I_m - x_{-j}^m\tilde{\beta}_{-j}^m\right) + \frac{1}{4n_m}x_j^{mT}x_j^m\beta_j^m + \tilde{\lambda}_{jm}\frac{\partial|\beta_j^m|}{\partial\beta_j^m} + A\beta_j^m + B = 0 \quad (14)$$

将式(14)整理可得式(15):

$$\left(\frac{1}{4n_m}x_j^{mT}x_j^m + A\right)\beta_j^m + \tilde{\lambda}_{jm}\frac{\partial|\beta_j^m|}{\partial\beta_j^m} = \frac{1}{4n_m}x_j^{mT}\left(z^m - \tilde{\beta}_0^m I_m - x_{-j}^m\tilde{\beta}_{-j}^m\right) - B \quad (15)$$

令 $\frac{1}{4n_m}x_j^{mT}\left(z^m - \tilde{\beta}_0^m I_m - x_{-j}^m\tilde{\beta}_{-j}^m\right) - B = D$，并对式(15)两边取绝对值，可得:

$$\left|\frac{1}{4n_m}x_j^{mT}x_j^m + A\right||\beta_j^m| + \tilde{\lambda}_{jm} = |D|$$

则 $|\beta_j^m| = \frac{(|D| - \tilde{\lambda}_{jm})_+}{\left|\frac{1}{4n_m}x_j^{mT}x_j^m + A\right|}$，本文发现 β_j^m 和 D 符号相同，所以:

$$\beta_j^m = \mathrm{sgn}(D)\frac{(|D| - \tilde{\lambda}_{jm})_+}{\frac{1}{4n_m}x_j^{mT}x_j^m + A} = \left[\frac{x_j^{mT}x_j^m}{4n_m} + \lambda_2\frac{M-1}{\tilde{\beta}_j^{m2} + \tau^2}\right]^{-1}S\left(D, \tilde{\lambda}_{jm}\right)$$

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其中 $D=\frac{1}{4n_m}x_j^{mT}\left(z^m-\tilde{\beta}_0^m I_m-x_{-j}^m\beta_{-j}^m\right)+\frac{\lambda_2}{\sqrt{\tilde{\beta}_j^{m2}+\tau^2}}\sum_{m\neq l}\frac{\tilde{\beta}_j^l}{\sqrt{\tilde{\beta}_j^{l2}+\tau^2}}$，$\tilde{\lambda}_{jm}=p'\left(\sum_{m=1}^{M}p(|\tilde{\beta}_j^m|;\lambda_1,a);\lambda_1,b\right)p'(|\tilde{\beta}_j^m|;\lambda_1,a)$ 和 $I_m=(1,1,\cdots,1)^T$ 是 $n_m\times 1$ 的列向量。

本文将上述求解过程应用于所有的 M 个数据集，并依次对 d 个变量的回归系数进行迭代，直至系数收敛为止，便可以得到最终的系数估计值。

算法可分为互相嵌套的两个部分，外循环负责更新 β_j，内循环负责计算在每一个数据集上的系数 β_j^m，具体算法整理如下:

算法(组坐标下降法)

1. 给定初始值 $\beta^{[0]}=(\beta_0^{[0]},\beta_1^{[0]}\cdots,\beta_d^{[0]})$ 和收敛精度，记已循环次数 $s=0$;
2. 根据循环初始值，由式(10)计算截距项 β_0 的估计值并更新;
3. 对每个 $j\in(1,\cdots,p)$，固定 $\beta_k^{[s]}(k\neq j)$，对 $\beta_j=(\beta_j^1,\cdots,\beta_j^M)^T$ 进行估计:
   (1)计算
   $\tilde{p}(x_i^m)=1/[1+\exp(-(\tilde{\beta}_0^m+\sum_{j=1}^{d}x_{ij}^m\tilde{\beta}_j^m))]$
   $z_i^m=\tilde{\beta}_0^m+\sum_{j=1}^{d}x_{ij}^m\tilde{\beta}_j^m+\{[y_i^m-\tilde{p}(x_i^m)]/\nu\}$
   $D=\frac{1}{4n_m}x_j^{mT}\left(z^m-\tilde{\beta}_0^m I_m-x_{-j}^m\beta_{-j}^m\right)+\frac{\lambda_2}{\sqrt{\tilde{\beta}_j^{m2}+\tau^2}}\sum_{m\neq l}\frac{\tilde{\beta}_j^l}{\sqrt{\tilde{\beta}_j^{l2}+\tau^2}}$
   $\tilde{\lambda}_{jm}=p'\left(\sum_{m=1}^{M}p(|\tilde{\beta}_j^m|;\lambda_1,a);\lambda_1,b\right)p'(|\tilde{\beta}_j^m|;\lambda_1,a)$
   (2)更新 β_j^m
   $\beta_j^{m[s+1]}=[x_j^{mT}x_j^m/(4n_m)+\lambda_2(M-1)/(\tilde{\beta}_j^{m2}+\tau^2)]^{-1}S(D,\tilde{\lambda}_{jm})$
4. 更新 s 为 $s+1$;
5. 重复步骤 2、3、4 直到收敛。

通过以上算法的循环迭代，可以得到的收敛值即为最终的系数估计值。

(三)调和参数选择

式(1)包含调和参数 λ_1 和 λ_2，MCP 惩罚的正则化参数 a 和 b，以及符号惩罚的参数 τ。参数的选取可能会直接影响模型的效果，因此对他们的恰当选取至关重要。对于 MCP 惩罚中的正则化参数 a，根据 Huang 等(2017)[14] 的研究，发现只要 a 不是特别大或者特别小，模型结果对 a 并不敏感，一般可以令 $a=6$。根据有关 MCP 双层变量惩罚的研究，令 $b=Ma\lambda_1^2/2$ 作为内外层惩罚的连接参数。符号惩罚中参数 τ 的作用是更好地近似符号惩罚并使其可导，只要这个参数值不是特别大，均能够满足上述条件，借鉴 Fang 等(2018)的设置，本文设 $\tau^2=0.5$。对于调和参数 λ_1 和 λ_2 的选择，使用的是测试集验证法，根据不同的 λ 建立一系列模型，使测试集 AUC 最大化来选择最优的 λ。在选择最优 λ 之前，需要先确定 λ 的范围，λ 达到最大值 $\lambda_{\max}$ 时惩罚力度最大，这时所有系数均为 0，从满足使系数全部为 0 的 λ 中选择一个最小的值作为 λ 的最大值 $\lambda_{\max}$。$\lambda_{\min}$ 通常取非常接近 0 的数，或者取 $\lambda_{\max}$ 的很小比例，如 $\lambda_{\min}=0.001\lambda_{\max}$，由此可确定参数的取值范围为 $[\lambda_{\min},\lambda_{\max}]$。

三、模拟分析

为了验证本文提出的模型的预测精度和变量选择效果，设计了三组具有代表性的模拟实验，每组模拟 100 次。关于模型比较，本文选择两种常用的方法: 一是合并模型(Pool Model)，即将多个数据集简单合并为一个数据集再进行建模; 二是独立模型(Meta Model)，即将多个数据集看作独立数

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

据集，对每个数据集分别独立建模。为了与本文所提出的整合模型具有可比性，这两个模型的变量选择惩罚函数都采用 MCP 函数。模型效果的评价标准主要从两个方面考虑，一是变量选择效果，即是否能准确选择出显著变量，评价指标主要有模型显著变量个数（P）、真实显著变量个数（TP）和伪发现率（FDR），其中 P 和 TP 越接近真实值越好，FDR 是将不显著变量识别为显著变量的错误率，这个指标越小越好；二是模型的预测效果，评价指标主要有样本外预测准确率（Accuracy）、真阳性率（TPR）和 AUC 值，这三个指标值越大越好。

设数据集数 $M = 4$，变量数 $d = 50$，每个数据集先生成样本数 $n = 200$，其中 $n = 100$ 用于训练模型，剩下的 $n = 100$ 作为测试集。参考 Liu 等（2014）[15] 的设置，模拟数据生成方式如下：①X 服从多元正态分布；②变量间的相关系数采用两种方案设置，一是自相关系数结构，即 $\mathrm{cov}(X_i, X_j) = \rho^{|i-j|}$，设 $\rho = 0.2$、0.7 分别表示弱相关和强相关；另一种为带状相关系数结构，又分为 $Band_1$ 和 $Band_2$ 两种情况：$Band_1$ 结构的设置为若 $|i-j| = 1$，则 $\mathrm{cov}(X_i, X_j) = 0.33$，否则 $\mathrm{cov}(X_i, X_j) = 0$；$Band_2$ 结构的设置为若 $|i-j| = 1$，$\mathrm{cov}(X_i, X_j) = 0.6$，若 $|i-j| = 2$，$\mathrm{cov}(X_i, X_j) = 0.33$，否则 $\mathrm{cov}(X_i, X_j) = 0$；③截距项 $\beta_0 = 0$；④被解释变量第 i 次的观测值 $y_i^m \sim Bernoulli(1, p_i^m)$，其中 $p_i^m = \Pr(y_i^m = 1 \mid X_i^m) = 1/[1 + \exp(-X_i^m \beta^m)]$；⑤关于系数的稀疏结构设置，考虑三种情况：变量稀疏结构完全相同、变量稀疏结构大部分相同和变量稀疏结构大部分不相同，具体详见模拟 1 ~ 3 的设置。

模拟 1：变量稀疏结构完成相同。四个数据集具有相同的变量稀疏结构，每个数据集均有 13 个显著变量，为 $(X_1^m, X_2^m, X_3^m, X_{11}^m, X_{12}^m, X_{13}^m, X_{21}^m, X_{22}^m, X_{23}^m, X_{31}^m, X_{32}^m, X_{33}^m, X_{34}^m)$，$m = 1, \cdots, M$，显著变量系数从均匀分布 $U(2,3)$ 中随机产生。

模拟 2：变量稀疏结构大部分相同。每个数据集也有 13 个显著变量，其中有 9 个相同的显著变量 $(X_1^m, X_2^m, X_3^m, X_{11}^m, X_{12}^m, X_{13}^m, X_{21}^m, X_{22}^m, X_{23}^m)$，$m = 1, \cdots, M$；各有 4 个不同的显著变量，分别为 $(X_{27}^1, X_{28}^1, X_{29}^1, X_{30}^1)$、$(X_{32}^2, X_{33}^2, X_{34}^2, X_{35}^2)$、$(X_{38}^3, X_{39}^3, X_{40}^3, X_{41}^3)$、$(X_{44}^4, X_{45}^4, X_{46}^4, X_{47}^4)$，显著变量的系数从均匀分布 $U(2,3)$ 中随机产生。

模拟 3：变量稀疏结构大部分不相同。每个数据集依然有 13 个显著变量，其中前 4 个解释变量为相同显著变量，即 $(X_1^m, X_2^m, X_3^m, X_4^m)$，$m = 1, \cdots, M$；各有 9 个不同的显著变量，分别是 $(X_7^1, X_8^1, X_9^1, X_{10}^1, X_{11}^1, X_{12}^1, X_{13}^1, X_{14}^1, X_{15}^1)$、$(X_{18}^2, X_{19}^2, X_{20}^2, X_{21}^2, X_{22}^2, X_{23}^2, X_{24}^2, X_{25}^2, X_{26}^2)$、$(X_{29}^3, X_{30}^3, X_{31}^3, X_{32}^3, X_{33}^3, X_{34}^3, X_{35}^3, X_{36}^3, X_{37}^3)$、$(X_{41}^4, X_{42}^4, X_{43}^4, X_{44}^4, X_{45}^4, X_{46}^4, X_{47}^4, X_{48}^4, X_{49}^4)$，显著变量的系数从均匀分布 $U(1,2)$ 中随机产生。

通过对模拟 1 ~ 3 的结果进行分析发现，总的来看，在预测效果方面，整合模型在三种模拟设置下的预测准确率、TPR 和 AUC 值基本上高于合并模型和独立模型，说明整合不同来源的数据可以提高模型的预测效果，而变量间不同的相关系数结构对模型的预测效果影响不大。在变量选择效果上，整合模型在不同数据集变量稀疏结构完全相同时能够选出所有的显著变量，而且从 FDR 看也明显优于合并模型和独立模型；在模拟 1 中，整合模型能选出更多的显著变量，其中 TP 也是最高的，而 FDR 往往也是最低的；在模拟 2 中，整合模型能选出更多的显著变量，并且 FDR 与合并模型相当，均要优于独立模型；在模拟 3 中整合模型的变量选择效果与合并模型相当，均优于独立模型。在模拟 1 ~ 3 中，在变量具有自相关系数结构下，随着自相关系数增加，整合模型变量选择 FDR 值增大，TP 值减小，说明变量间的自相关系数会影响模型变量的选择效果。综上所述，整合模型在预测效果和变量选择效果上相对于合并模型与独立模型有更好的表现。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1　　模拟 1 结果

相关系数	模型	P	TP	FDR	Accuracy	TPR	AUC
$\rho = 0.2$	整合	54.46	52.00	0.0320	0.9384	0.9386	0.9852
	合并	53.43	51.01	0.0428	0.9066	0.9093	0.9668
	独立	56.59	46.62	0.1743	0.8436	0.8345	0.9167
$\rho = 0.7$	整合	57.05	51.96	0.0724	0.9549	0.9548	0.9917
	合并	55.68	47.76	0.1353	0.9233	0.9208	0.9780
	独立	38.46	29.58	0.2280	0.8758	0.8801	0.9434
$Band_1$	整合	53.35	52.00	0.0216	0.9413	0.9410	0.9861
	合并	52.82	50.80	0.0364	0.9086	0.9080	0.9688
	独立	52.53	43.55	0.1701	0.8441	0.8394	0.9171
$Band_2$	整合	54.97	51.96	0.0476	0.9512	0.9508	0.9903
	合并	53.67	49.53	0.0746	0.9210	0.9163	0.9759
	独立	41.74	34.72	0.1665	0.8649	0.8610	0.9342

表 2　　模拟 2 结果

相关系数	模型	P	TP	FDR	Accuracy	TPR	AUC
$\rho = 0.2$	整合	58.13	47.99	0.1558	0.8665	0.8619	0.9383
	合并	51.71	45.61	0.1027	0.8530	0.8533	0.9255
	独立	53.69	43.43	0.1885	0.8225	0.8322	0.8966
$\rho = 0.7$	整合	55.08	45.86	0.1545	0.9007	0.9002	0.9640
	合并	56.75	43.08	0.2070	0.8915	0.8905	0.9574
	独立	37.74	28.36	0.2452	0.8609	0.8633	0.9306
$Band_1$	整合	56.17	47.9	0.1378	0.8707	0.8696	0.9416
	合并	50.54	45.46	0.0913	0.8574	0.8575	0.9307
	独立	50.91	41.05	0.1925	0.8247	0.8200	0.8981
$Band_2$	整合	52.78	46.51	0.1105	0.8925	0.8936	0.9585
	合并	51.56	43.95	0.1161	0.8847	0.8851	0.9524
	独立	40.84	33.11	0.1872	0.8498	0.8444	0.9213

表 3　　模拟 3 结果

相关系数	模型	P	TP	FDR	Accuracy	TPR	AUC
$\rho = 0.2$	整合	64.92	45.04	0.2711	0.8369	0.8484	0.9112
	合并	65.80	46.36	0.2446	0.8323	0.8299	0.9078
	独立	53.02	43.52	0.1767	0.8217	0.8234	0.8946
$\rho = 0.7$	整合	47.70	40.68	0.1394	0.9048	0.9078	0.9668
	合并	55.78	42.60	0.1661	0.9119	0.9132	0.9695
	独立	34.78	28.58	0.1753	0.8836	0.8816	0.9495
$Band_1$	整合	56.06	44.96	0.1711	0.8483	0.8479	0.9241
	合并	56.40	44.72	0.1744	0.8455	0.8446	0.9221
	独立	50.16	41.28	0.1755	0.8312	0.8288	0.9055
$Band_2$	整合	53.20	42.20	0.1437	0.8850	0.8798	0.9541
	合并	52.18	44.12	0.1348	0.8903	0.8801	0.9573
	独立	35.52	28.20	0.2033	0.8576	0.8669	0.9298

四、实证分析

传统的个人信用评分模型往往把城乡数据合并在一起构建统一的评分模型，但是由于我国经济发展呈现城乡二元结构特征，城乡经济、文化、居民富裕程度、居住环境等方面的差异，导致城乡居民在消费观念上存在较大的区别，观念的差别在一定程度上影响着城乡居民的消费行为。因此，用统一的信用评分模型去评估城乡居民的信用不符合我国的国情，忽略了城乡居民在信用行为上的差异性; 另一方面，如果针对城乡数据集单独建立模型，会忽略数据集间的关联性，比如出现变量稀疏结构差异大、模型系数符号相反等，从而影响模型的解释性和预测精度等。因此，本文将城乡

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

居民的征信数据集看作两个不同来源的数据集，利用本文提出的整合模型，同时考虑了城乡数据集间的关联性和异质性，并能同时对模型进行系数估计和变量选择，自动筛选出每个数据集的显著变量，且能鼓励同一变量在不同数据集的系数符号具有相似性。

数据来自于某商业银行的信用卡客户资料库，分城市和农村两个数据集。数据变量共有 21 个，主要涉及个人基本情况、家庭基本情况、收入情况、信用卡使用情况、不良记录等。变量说明详见表 4。

表 4　　变量说明

变量	说明	变量	说明
违约	1 为违约客户；0 为非违约客户	学历	1 表示小学及以下；0 表示其他
逾期是否 30 天	1 表示是；0 表示否		1 表示初中；0 表示其他
是否有呆账记录	1 表示是；0 表示否		1 表示高中；0 表示其他
借款余额大于 15 万元	1 表示是；0 表示否		1 表示专科；0 表示其他
退票记录	1 表示有退票记录；0 表示没有	职业	1 表示有职业；0 表示无职业
拒往记录	1 表示有拒往记录；0 表示没有	房产	1 表示父母所有；0 表示其他
他行强制停卡记录	1 表示有他行强制停卡记录；0 表示没有		1 表示本人所有；0 表示其他
信用卡张数	取值为 [1，5]，其中 5 表示 5 张及 5 张以上		1 表示配偶所有；0 表示其他
使用频率	0 表示没有用；1 表示很少用；2 表示偶尔用；3 表示经常用；4 表示天天用		1 表示住宿舍；0 表示其他
户籍所在地域	1 表示北部区域；0 表示其他		1 表示租赁；0 表示其他
	1 表示中部区域；0 表示其他	个人平均月收入	取值 $[0,+\infty)$
	1 表示南部区域；0 表示其他。	个人平均月开销	取值 $[0,+\infty)$
性别	1 表示男；0 表示女	家庭平均月收入	取值 $[0,+\infty)$
年龄	$X_3 \in [15,60]$；$X_3 \in N$	月信用卡刷卡金额	取值 $[0,+\infty)$
婚姻状况	1 表示未婚；0 表示其他	家庭人口数	取值为整数

对连续变量进行标准化处理，定性变量进行虚拟化处理。由于数据存在不平衡问题，参考方匡南等(2014)，对于不平衡问题采用“减少多数法”处理，将农村 11590 笔数据全部用于建模，同时运用抽样技术从城市数据集中抽取 11590 条数据用于建立模型。在对数据集进行处理之后两个数据集共得到样本 23180 个，违约客户与非违约客户之比约为 3∶7，随机选择 70% 用于训练模型，剩余 30% 作为测试集验证模型的预测性能和稳健性。为了减少随机性对模型产生的影响，本文对上述过程重复操作 50 次，分别得到 50 个训练集和测试集。从 50 次预测结果的平均值来看，整合模型的预测准确率(Accuracy)最高，独立模型的预测准确率最低，三个模型的 TPR 值均为 1，说明均能够很好地识别违约客户，能尽可能减少违约客户给银行带来的风险，整合模型的 AUC 值也是最高的。所以综合来看，整合模型的预测效果要好于传统的合并模型和独立模型。

从变量选择结果看，两个数据集都显著的变量主要有逾期是否 30 天、他行强制停卡记录、房产变量、婚姻情况、学历(高中)、月均刷卡额等，这与王小燕等(2014)的结论基本一致。其中逾期和他行停卡均属于过往的不良信用记录，以往的记录反映了客户整体的信用情况，有不良信用记录的客户更容易违约；对于房产，无论是自己、配偶还是父母有房都属于比较稳定的情况，相对于没有房的客户信用程度更高；未婚客户相对不够稳定，没有家庭的负担，可能消费支出缺乏有效规划和管理，违约风险更高；而高学历人群的违约风险相对较小；每月刷卡额度越高，即每月消费支出越高，违约风险也就越高。

此外，户籍地区在北部和中部的这两个变量在城市数据集不显著，而在农村数据集的系数为负，说明这两个区域的农村相对于其他区域来说信用程度较高；年龄变量在城市数据集中系数为负，但在农村数据集不显著，可能的原因是城市居民主要以工资收入为主，随着年龄增长，工作也更稳定，收入也在增长，自身财力增加，违约可能性减小，而农村的收入更多地取决于农产品，与工作年限或年龄等关系不大。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

五、总结

本文所研究的主要是针对来源不同、样本不同、变量相同的数据集进行建模分析,即“样本”方向的整合建模,提出了基于多源数据融合的 logistic 回归模型,通过损失函数综合不同来源的数据集,从统计角度考虑数据集的异质性,并在损失函数上添加变量选择惩罚函数和系数符号相似惩罚函数,即考虑高维数据的变量选择,一方面要选出每个数据集上的显著变量,另一方面又要考虑不同数据集间的关联性。通过模拟发现,整合模型在变量选择和分类效果方面都具有优势。此外,将本文的模型应用于城市和农村两个数据集的个人信用评分中,通过某大型商业银行信用卡数据实证分析,检验整合模型在实际应用中的效果。研究结果显示,整合模型的样本外预测准确率和 AUC 值均高于合并数据集模型和独立数据集模型,在实际应用中有很好的表现。

综上所述,整合模型融合了多个数据源,允许不同数据集在同一变量上的系数有所不同,但鼓励系数符号在不同数据集具有相似性;提高了模型估计精度,为分析个人信用问题提供了新的解决方案。需要说明的是,本文所提出的模型虽然主要应用于城市和农村两个数据集的个人信用评分中,但是该模型很容易扩展到多个地区的个人信用评分模型中,也可以应用到其他多源数据融合的分类建模问题中。

参考文献

[1] 胡心瀚,叶五一,缪柏其. 上市公司信用风险分析模型中的变量选择[J]. 数理统计与管理,2012,31(6):1117-1124.

[2] 方匡南,章贵军,张惠颖. 基于 Lasso-logistic 模型的个人信用风险预警方法[J]. 数量经济技术经济研究,2014,(2):125-136.

[3] Wiginton J C. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior[J]. Journal of Financial & Quantitative Analysis,1980,15(3):757-770.

[4] West D. Neutral Network Credit Scoring Models[J]. Computers & Operations Research,2000(27):1131-1152.

[5] 李志辉,李萌. 我国商业银行信用风险识别模型及其实证研究[J]. 经济科学,2005(5):61-71.

[6] 王大荣,张忠占. 线性回归模型中变量选择方法综述[J]. 数理统计与管理,2012,29(4):615-627.

[7] 王小燕,方匡南,谢邦昌,等. Logistic 回归的双层变量选择研究[J]. 统计研究,2014,31(9):107-112.

[8] 胡小宁,何晓群,马学俊. 基于 Group MCP Logistic 模型的个人信用评价分析[J]. 现代管理科学,2015(8):18-20.

[9] 方匡南,范新妍,马双鸽. 基于网络结构 Logistic 模型的企业信用风险预警[J]. 统计研究,2016,33(4):50-55

[10] 马双鸽,王小燕,方匡南. 大数据的整合分析方法[J]. 统计研究,2015,32(11):3-11.

[11] Breheny P, Huang J. Group Descent Algorithms for Nonconvex Penalized Linear and Logistic Regression Models with Grouped Predictors[J]. Statistics & Computing,2015,25(2):173-187.

[12] Fang K, et al. Integrative Sparse Principal Component Analysis[J]. Journal of Multivariate Analysis,2018(166):1-16.

[13] Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables[J]. Journal of the Royal Statistical Society,2006,68(1):49-67.

[14] Huang Y, et al. Promoting Similarity of Sparsity Structures in Integrative Analysis with Penalization[J]. Journal of the American Statistical Association,2017,112(517):342-350.

[15] Liu J, Ma S, Huang J. Integrative Analysis of Cancer Diagnosis Studies with Composite Penalization[J]. Scandinavian Journal of Statistics,2014,41(1):87-103.

作者简介

方匡南,男,2010 年毕业于厦门大学计划统计系,获经济学博士学位,现为厦门大学经济学院教授、博士生导师。研究方向为数据挖掘、应用统计、金融风险管理。

赵梦恋,女,2018 年毕业于厦门大学经济学院,获经济学硕士学位,现为深圳前海微众银行股份有限公司研究员。研究方向为数据挖掘。

(责任编辑: 郭明英)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net