E-Commerce Letters 电子商务评论, 2025, 14(1), 1354-1360
Published Online January 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.141168

Hans汉斯

基于稀疏逻辑回归的信用风险评估模型

王丽华，彭定涛*

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2024年10月17日；录用日期：2024年11月4日；发布日期：2025年1月14日

摘 要

随着经济的持续增长和金融科技的不断发展，个人信贷作为一种满足消费需求的金融工具，其市场规模自然随之扩大。受到经济下行压力、不良贷款行为增加与各种突发变故的影响，个人信贷违约率逐渐上升，一个完善且高效的个人信用评估模型其重要性不言而喻。在信用评估过程中，通过一系列的具体指标和因素去判断个人的信用风险，在庞大的市场规模下，需要巨量的资源投入。本文提出了一种基于稀疏优化的逻辑回归模型，其能在保持一定准确度的情况下快速地得出个人风险评估结果。最后通过真实数据，验证所提出稀疏逻辑回归模型的有效性。

关键词

信用风险，稀疏优化，逻辑回归

Credit Risk Assessment Model Based on Sparse Logistic Regression

Lihua Wang, Dingtao Peng*

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Oct. 17th, 2024; accepted: Nov. 4th, 2024; published: Jan. 14th, 2025

Abstract

With the continuous growth of the economy and the development of financial technology, the market scale of personal credit, as a financial tool to satisfy consumer demand, has naturally expanded. Influenced by the economic downward pressure, the increase of non-performing loan behaviors and various unexpected changes, the default rate of personal credit is gradually rising, and the importance of a perfect and efficient personal credit assessment model is self-evident. In the process of credit

*通讯作者。

文章引用: 王丽华, 彭定涛. 基于稀疏逻辑回归的信用风险评估模型[J]. 电子商务评论, 2025, 14(1): 1354-1360.
DOI: 10.12677/ecl.2025.141168

assessment, a series of specific indicators and factors are used to judge the credit risk of an individual, which requires a huge amount of resources under a huge market scale. In this paper, a logistic regression model based on sparse optimization is proposed, which can quickly produce individual risk assessment results while maintaining a certain degree of accuracy. Finally, the effectiveness of the proposed sparse logistic regression model is verified by real data.

Keywords

Credit Risk, Sparse Optimization, Logistic Regression

1. 引言

近年来，随着科技与社会经济的飞速发展，大数据技术也广泛应用于各行各业，例如房地产，金融，电子商务等。依托于互联网用户的逐年上升，互联网金融等业务也越发繁荣。由于其快速便捷等优势，“网上银行”也是备受青睐。随着普惠金融和金融大数据的不断发展，随着大众消费模式的转变，信贷消费日益成为消费支出的主导方式，消费贷款规模逐年扩大。一方面个人信贷业务的发展刺激了消费，推动了经济发展。另一方面信贷业务也伴随着隐形风险，个人信贷违约率逐渐上升。因此，信用风险评估始终是金融研究的核心主题之一。

随着消费贷款规模的逐渐扩大和信用风险的不断增加，对其进行防范和控制刻不容缓。而对信贷机构而言，其核心在于如何构建一个合理可靠的信贷风险模型以期评估信贷申请人的信用风险。信用风险评估也从最初仅依靠专家主观经验判断，到逐渐客观的计量模型方法，再到现在的人工智能模型，可见针对信用风险的评估也一直在不断改进。2014 年，方匡南等人将 Lasso-Logistic 模型引入个人信用评估并用实证分析证明该方法的有效性[1]；胡越提出基于正则化下的支持向量机对信用风险进行评估[2]；许迩璇选用随机森林模型对个人信贷风险进行研究，并与 Logistic 回归模型进行对比，证明了其模型对风险评估的有效性[3]。针对信用风险评估的研究和方法已然取得一定成效。

然而，伴随着大数据时代的到来，含有海量信息和特征的信贷数据集对所需评估模型提出了更高的要求。例如年龄、职业、国籍、学历等诸多因素对个人信用风险评估结果都可能会产生或多或少的影响。一方面，变量的多重共线性对模型的可解释性与预测准确性会有影响；另一方面，一些无关自变量的选入也会增加模型的复杂度。

本文将近年来在机器学习信号处理，模式识别中的研究热点——“稀疏性”引入到信用风险评估模型中，提出基于稀疏优化的逻辑回归模型。稀疏优化问题最早由 David Donoho 等人于 1998 年提出[4]，旨在通过解的稀疏性结构来构建数学模型，以提高模型解的稳定性与鲁棒性。2005 年，Candes 等人[5]给出该模型的数学理论，并为稀疏优化模型奠定了理论根基。近年来，稀疏优化理论在机器学习，基因选择，投资组合等众多领域受到广泛应用。利用稀疏性不仅能对向量进行充分地压缩，从而节约储存空间，还可以从海量的数据中挖掘出有价值的信息让复杂问题得以简化。

本文脉络如下，首先在第二节介绍个人信用风险评估的意义与稀疏逻辑回归模型，并给出求解该模型的邻近梯度算法。在第三节中，将应用所提算法模型求解真实数据。

2. 个人信用评估与逻辑回归模型

2.1. 个人信用评估

信用评估是金融机构利用大数据分析收集的大量信息，并通过收集到的数据信息对企业和个人的信用状况进行综合评估，通过分析个人信用行为历史和金融机构的商业信用行为数据得出的个人信用评估模型。利用数据挖掘等技术得到个人信用评估模型，可以帮助我们更加准确有效地预测和分析个人商业信用行为风险表现。一方面可以提高操作的准确性和效率，另一方面可有效降低金融机构与信贷提供相关的成本。因此信用评估是现代金融机构信用评估框架中不可或缺的工具。

在银行信贷业务中，信用评估常用于评估借款人违约风险。其中违约风险即贷款申请人到期无法偿还贷款的风险。而通过收集并系统分析用户的人口特征、信用历史、行为记录、交易日志等大量历史数据，可以挖掘数据中潜在行为模式和信用属性，从而开发出旨在评估用户信誉的信用评估模型。该信用评分模型可以从个人信用文件中提取各种属性，并据此评估用户的信用状况。通过信用评分模型可评估相关风险并相应地调整信用评分，同时贷款申请人也可以得到及时的反馈，进而大大提高信贷决策的效率。

从本质上讲，信用风险评估问题可以被建模为一个二分类问题，即根据借贷人是否有可能违约，区分可正常还款的好客户和有违约风险的坏客户。那么，如何对客户进行准确有效的分类，并在拓展信贷业务的同时有效防控信用风险，也是当前亟待解决的重要问题。因此，建立有效的信用风险评估模型，以应对可能出现的信用风险具有重要的现实意义。

2.2. 稀疏逻辑回归问题

基于变量的稀疏性，通常会借助向量的 ℓ_0 范数来刻画，即以下 ℓ_0 正则优化模型[6]：

$$\min_{x\in\mathbb{R}^n} \quad f(x)+\lambda\|x\|_0$$

其中 $f(x)$ 为损失函数，$\lambda>0$ 是正则化参数，$\|x\|_0$：所有非零元素的个数。由于 ℓ_0 范数非凸且非光滑，因此该问题是一个 NP 难的。于是研究者们考虑了 ℓ_0 范数的最紧凸松弛——ℓ_1 范数，即如下模型[7]：

$$\min_{x\in\mathbb{R}^n} \quad f(x)+\lambda\|x\|_1$$

但 ℓ_1 范数和同样是非光滑的，且大量实证研究表明 ℓ_1 正则优化模型解的稳健性与稀疏诱导性都比较差，Wang 等人[8]从理论上证明了非凸的 ℓ_q $(0<q<1)$正则项比凸的 ℓ_1 正则项更能诱导解的稀疏性，非凸正则项 ℓ_q $(0<q<1)$范数定义如下：

$$\|x\|_q^q := \sum_{i=1}^{n}|x_i|^q,$$

ℓ_q 正则优化模型[8]-[11]为：

$$\min_{x\in\mathbb{R}^n} \quad f(x)+\lambda\|x\|_q^q,$$

当逻辑函数作为损失函数时，本文研究如下稀疏逻辑回归问题：

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{m}\sum_{i=1}^{m}\left(\ln\left(1+\mathrm{e}^{\langle a_i,x\rangle}\right)+b_i\langle a_i,x\rangle\right)+\lambda\|x\|_q^q. \tag{1}$$

其中 $a_i\in\mathbb{R}^n$，$b\in\mathbb{R}^m$，$\lambda>0$。

2.3. 邻近梯度算法求解稀疏逻辑回归问题

下面，记目标函数为：

$$F(x) := f(x) + \lambda \|x\|_q^q,$$

其中

$$f(x) := \frac{1}{m}\sum_{i=1}^{m}\left(\ln\left(1+e^{\langle a_i, x\rangle}\right) + b_i \langle a_i, x\rangle\right),$$

由于损失函数 f 是光滑的凸函数，考虑用邻近梯度算法求解问题(1)。我们首先给出邻近算子的定义。

设 $f:\mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ 为一正常下半连续函数，给定 $v\in\mathbb{R}^n$ 和 $\lambda>0$，算子 $Prox_{\lambda f}$ 定义为：

$$Prox_{\lambda f}(v) := \arg\min\left\{\lambda f(x) + \frac{1}{2}\|x-v\|^2 : x\in\mathbb{R}^n\right\},$$

由邻近算子的定义，可以直接给出 ℓ_q 范数的算子[9] [Proposition 2.3]

对给定 $\lambda>0$ 且 $0\le q<1$，令

$$c(\lambda,q) = \left(2\lambda^{1-q}\right)^{\frac{1}{2-q}},\quad \kappa(\lambda,q) := (2-q)\lambda^{\frac{1}{2-q}}\left(2(1-q)\right)^{\frac{q+1}{q-2}},$$

引理 2.2.1 [9]给定 $x\in\mathbb{R}$，$\lambda>0$，$q\in[0,1)$，$|\cdot|^q$ 的邻近算子为：

$$Prox_{\lambda|\cdot|^q}(z) := \arg\min_x\left\{\lambda|x|^q + \frac{1}{2}\|x-z\|^2\right\} =: \varphi(x)$$

$$= \begin{cases} \{0\}, & |z| < \kappa(\lambda,q), \\ \{0, \operatorname{sgn}(z)c(\lambda,q)\}, & |z| = \kappa(\lambda,q), \\ \{\operatorname{sgn}(z)\varpi_q(|z|)\}, & |z| > \kappa(\lambda,q), \end{cases} \tag{2}$$

其中 $\varpi_q(z) := \{x : z - x + \lambda q \operatorname{sgn}(x)x^{q-1}, x>0\}$ 是唯一的。此外，若 $Prox_{\lambda|\cdot|^q}(z)$ 包含一个非零点 x^*，则

$$|x^*| \ge c(\lambda,q),\quad \varphi''(|x^*|) \ge 1-\frac{q}{2} > \frac{1}{2}。$$

由于损失函数是光滑的，对上述逻辑回归问题，提出邻近梯度算法进行求解：

邻近梯度算法(PGM)：

步 1：初始化 x^0，$\alpha_0>0$，$k=0$，$\lambda>0$。

步 2：While 未达到收敛准则 do

$$z^k = x^k - \alpha_k \nabla f(x^k),$$
$$w^k = Prox_{\lambda\alpha_k\|\cdot\|_q^q}(z^k),。$$

步 3：end。

3. 实例验证

由于国内对于个人隐私的注重，本文研究数据取自 UCI 数据库，利用邻近梯度算法求解稀疏逻辑回归问题对个人客户信贷信用进行评估。数据采用由 Hans Hofmann 教授整理的德国信贷数据集，该数据集被广泛应用于个人信用评估方法的验证，总共包含 1000 份客户资料，其中每位客户包含 20 条属性，其中有 7 个为连续变量、13 个为离散变量，并给出了信用的好或坏的标注[12]。好客户用数字“2”表示，坏客户用数字“1”表示。具体说明如表 1 所示。

本文算法在 Matlab2019 和 Windows11 处理器为 Intel(R) Core(TM) i5-9500 CPU @3.00GHz 1.00GHz 上运行。为了便于计算，我们用数值来表示数据中的类别属性，用整数进行编码，如第一个属性的类别字符串为“A12”，“A1”表示该属性为第一个属性，“2”表示属性值为第二个类别。因此，采用数字

“2”编码字符串“A12”。我们将数据集分为训练组和测试组，每组有 500 个数据，并对输入样本进行归一化处理后利用算法进行计算。具体样本分布如表 2 所示。

Table 1. Description of customer attributes in the German credit database [12]
表 1. 德国信用数据库中客户属性说明[12]

属性名称	数值类型	取值范围
属性 1： 现有支票帐户的状态	字符	A11：<0 DM，A12：0 ≤ x < 200 DM，A13：≥200 DM/至少一年的薪水分配，A14：无支票帐户
属性 2： 期数或贷款持续月份	数字	[4, 72]
属性 3： 历史信用记录	字符	A30：未提取任何信用/已全额偿还所有信用额，A31：已偿还该银行的所有信用额，A32：已到期已偿还的现有信用额，A33：过去的还款延迟，A34：关键帐户/其他信用额现有(不在此银行)
属性 4： 借款目的	字符	A40：汽车(新)，A41：汽车(二手)，A42：家具/设备，A43：广播/电视，A44：家用电器，A45：修理，A46：教育，A47：(假期)，A48：再培训，A49：商业，A410：其他
属性 5： 借款额度	数字	[250, 18,424]
属性 6： 储蓄账户状态	字符	A61：... < 100，A62：100 ≤ ... < 500，A63：500 ≤ ... ≤ 1000，A64：... ≥ 1000，A65：未知/没有储蓄账户
属性 7： 当前就业状态	字符	A71：失业，A72：... < 1 年，A73：1 ≤ ... < 4 年，A74：4 ≤ ... < 7 年，A75：... ≥ 7 年
属性 8： 分期付款占可支配收入的百分比	数字	[1, 4]
属性 9： 性别与婚姻状态	字符	A91：男：离婚/分居，A92：女：已婚/丧偶，A93：男：单，A94：男：已婚/丧偶，A95：女：单
属性 10： 其他担保人	字符	A101：无，A102：共同申请人，A103：担保人
属性 11： 现居住地	数字	[1, 4]
属性 12： 财产状况	字符	A121：房地产，A122：建房协会储蓄协议/人寿保险，A123：汽车或其他，A124：未知/无属性
属性 13： 年龄	数字	[19, 75]
属性 14： 其他分期情况	字符	A141：银行，A142：商店，A143：无
属性 15： 房产状态	字符	A151：租房，A152：自己的，A153：免租房
属性 16： 信用卡数量	数字	[1, 4]
属性 17： 工作状态	字符	A171：失业/非技术人员－非居民，A172：非熟练－居民，A173：熟练员工/官员，A174：管理/个体经营/高素质的员工/官员
属性 18： 赡养人数	数字	[1, 2]
属性 19： 电话注册情况	字符	A191：无，A92：是的，以用户名已注册
属性 20： 是否有国外经历	字符	A201：有，A202：没有

Table 2. Sample distribution
表 2. 样本分布

	样本数量	履约样本数	履约样本所占比例(%)	违约样本数	违约样本所占比例(%)
训练样本	500	350	70	150	30
测试样本	500	350	70	150	30
全体样本	1000	700	70	300	30

分别选取 q = [0, 1/2, 2/3]三种情况下进行实验，其中实验参数选取如下：$\lambda = 2e-3$；$\alpha_0 = 30$；$\alpha_{k+1} = \alpha_k * rate$；$rate = 0.1$。三种 q 值下的逻辑回归系数如表 3 所示。

Table 3. Logistic regression coefficient
表 3. 逻辑回归系数

	q = 0	q = 1/2	q = 2/3
x1	−0.8891	−0.9209	−0.9233
x2	0.6675	0.6918	0.6885
x3	−0.6728	−0.6899	−0.6557
x4	0	0	0.0345
x5	0.7237	0.7452	0.7506
x6	−0.3236	−0.2890	−0.2974
x7	−0.4942	−0.4881	−0.4719
x8	0.5392	0.5520	0.5621
x9	0	−0.5557	−0.5172
x10	0	−0.2955	−0.2875
x11	0	−0.0563	−0.0760
x12	0.2041	0.0872	0.1583
x13	0	0	0
x14	−0.1560	−0.1870	−0.1974
x15	0	0	−0.2512
x16	0.3673	0.4369	0.4306
x17	0	0.1282	0.1499
x18	0	0.1295	0.1492
x19	0	0	0
x20	0	0	0

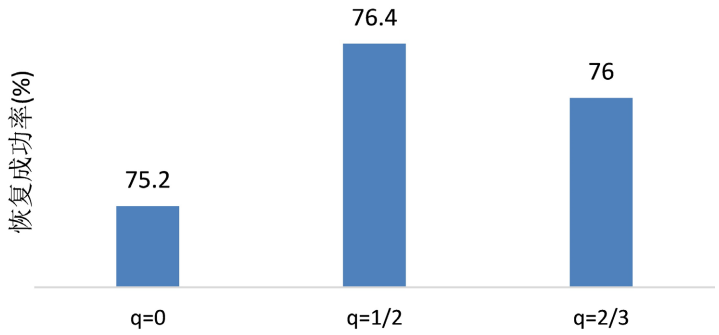

Figure 1. Model prediction accuracy under different q values
图 1. 不同 q 值下的模型预测正确率

同样，三种 q 值下的实验结果的准确率如图 1 所示。

根据上述实验结果，可以看到我们的算法对数据集可以进行有效预测，并且预测正确率也是较高的。

4. 总结与展望

本文提出稀疏逻辑回归模型，并用邻近梯度算法进行求解，对德国信贷数据集进行试验，实验结果证明，所提算法模型可以根据用户数据特征对用户信用进行有效预测。上述结果表明所提稀疏逻辑回归模型可以对客户信用进行有效评估预测，有助于降低个人信用风险。但是在预测的准确率上可以看到我们还有进步的空间，因此，在今后的研究工作中将继续对该模型进行调整以改善模型的预测正确率。

基金项目

国家自然科学基金(12261020)，贵州省科技计划(黔科合基础 ZK[2021]一般 009)和贵州省高层次留学人才创新创业择优资助重点项目([2018]03)。

参考文献

[1] 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, 31(2): 125-136.

[2] 胡越. 正则化下支持向量机的信用风险评估[D]: [硕士毕业论文]. 上海: 上海师范大学, 2017.

[3] 许迩璇. 基于随机森林模型的个人信贷风险研究[J]. 审计与理财, 2024(9): 55-58.

[4] Chen, S.S., Donoho, D.L. and Saunders, M.A. (2001) Atomic Decomposition by Basis Pursuit. *SIAM Review*, **43**, 129-159. https://doi.org/10.1137/s003614450037906x

[5] Candes, E.J. and Tao, T. (2005) Decoding by Linear Programming. *IEEE Transactions on Information Theory*, **51**, 4203-4215. https://doi.org/10.1109/tit.2005.858979

[6] Natarajan, B.K. (1995) Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, **24**, 227-234. https://doi.org/10.1137/s0097539792240406

[7] Chartrand, R. and Staneva, V. (2008) Restricted Isometry Properties and Nonconvex Compressive Sensing. *Inverse Problems*, **24**, Article 035020. https://doi.org/10.1088/0266-5611/24/3/035020

[8] Wang, H., Zhang, F., Shi, Y. and Hu, Y. (2021) Nonconvex and Nonsmooth Sparse Optimization via Adaptively Iterative Reweighted Methods. *Journal of Global Optimization*, **81**, 717-748. https://doi.org/10.1007/s10898-021-01093-0

[9] Zhou, S., Xiu, X., Wang, Y., *et al*. (2023) Revisiting Lq ($0 \le q < 1$) Norm Regularized Optimization. arXiv: 2306.14394.

[10] Peng, D., Xiu, N. and Yu, J. (2017) $S_{1/2}$ Regularization Methods and Fixed Point Algorithms for Affine Rank Minimization Problems. *Computational Optimization and Applications*, **67**, 543-569. https://doi.org/10.1007/s10589-017-9898-5

[11] Peng, D., Xiu, N. and Yu, J. (2018) Global Optimality Condition and Fixed Point Continuation Algorithm for Non-Lipschitz ℓ_p Regularized Matrix Minimization. *Science China Mathematics*, **61**, 1139-1152. https://doi.org/10.1007/s11425-016-9107-y

[12] 张岗岗. 稀疏组 Lasso 方法在个人信贷风险评估中的应用[D]: [硕士学位论文]. 济南: 山东大学, 2018.