

基于斜坡损失最小二乘几何NHSVM的信用卡客户违约概率预测分析

万婷婷

贵州大学数学与统计学院, 贵州 贵阳

收稿日期: 2024年7月7日; 录用日期: 2024年11月6日; 发布日期: 2024年11月13日

摘要

本文基于UCI机器学习库中的信用卡客户违约的数据, 以客户是否存在信用卡违约行为作为响应变量, 以23个描述客户信息以及客户每月的还款情况和还款金额的变量作为解释变量建立预测模型。为了提升金融服务领域对信用卡违约概率评估的准确性和效率, 本文提出了一种优化的预测模型——改进的斜坡损失最小二乘几何非平行超平面支持向量机(RLS-GNHSVM)。RLS-GNHSVM模型融合了斜坡损失函数和最小二乘几何非平行超平面支持向量机的优势, 旨在克服传统的凸损失函数对异常值敏感而导致性能不佳的问题。该模型不仅能够在数据含噪或存在异常值的情况下保持稳定的预测性能, 还显著优化了预测精度。在实证应用中, RLS-GNHSVM模型相较于其他三种主流模型, 在预测信用卡客户违约概率方面展现出了更高的效能和适用性, 为金融机构提供了更为精准的风险评估工具。

关键词

斜坡损失函数, 最小二乘, 非平行支持向量机, 信用卡客户违约, 预测模型

Analysis of Credit Card Customer Default Probability Prediction Using Least Squares Ramp Loss Geometric NHSVM

Tingting Wan

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Jul. 7th, 2024; accepted: Nov. 6th, 2024; published: Nov. 13th, 2024

Abstract

Based on the credit card customer default dataset from the UCI Machine Learning Repository, this

文章引用: 万婷婷. 基于斜坡损失最小二乘几何 NHSVM 的信用卡客户违约概率预测分析[J]. 电子商务评论, 2024, 13(4): 2756-2766. DOI: 10.12677/ec.2024.1341454

paper establishes a predictive model with the presence of credit card default behavior as the response variable and 23 explanatory variables detailing customer information, along with monthly repayment status and amounts during the data collection period. To enhance the accuracy and efficiency of credit card default probability assessment in the financial services sector, we propose an optimized predictive model known as the Refined Least Squares Ramp Loss Geometric Non-Parallel Hyperplane Support Vector Machine (RLS-GNHSVM). The RLS-GNHSVM model seamlessly combines the strengths of the Least Squares Ramp Loss function and the Geometric Non-Parallel Hyperplane SVM. This integration aims to address the shortcomings of traditional convex loss functions, which are prone to performance degradation due to sensitivity to outliers. The RLS-GNHSVM model not only maintains stable predictive performance amidst noisy data or the presence of outliers but also significantly enhances prediction accuracy. In empirical applications, the RLS-GNHSVM model demonstrates superior performance and applicability compared to three other mainstream models in predicting credit card customer default probabilities. By offering a more precise risk assessment tool, this model provides financial institutions with a powerful means to refine their decision-making processes and enhance their risk management capabilities.

Keywords

Ramp Loss, Least Squares, Non-Parallel SVM, Credit Card Customer Default, Predictive Models

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

21 世纪全球经济中，金融市场繁荣稳定成为经济增长和社会进步的关键。全球化与数字化推动金融市场复杂化、互联化，金融创新与服务模式变革加速。信用风险管理核心地位凸显，准确预测违约概率关乎金融机构安全、利润及市场稳定、投资者信心。消费金融、互联网金融普及带来便利，也伴随信用违约风险。传统信用评分系统面临市场与消费者行为变化挑战，难以全面捕捉新风险。为了应对这些挑战，金融机构和学术界开始探索更加先进、智能的信用风险评估方法。其中，基于大数据和机器学习技术的信用客户违约概率预测模型备受关注。这些模型能够处理海量、多维度的数据资源，通过复杂的算法和模型构建过程，挖掘出隐藏在数据背后的深层次规律，从而实现对客户违约风险的精准预测。

Vapnik 和 Cortes 于 20 世纪末提出了一种新的用于分类问题的工具[1]：支持向量机(Support vector machine, SVM)是用于监督学习的计算功能强大的工具，在各种现实世界的应用中成功且广泛地用于分类和回归。近年来，支持向量机的一个分支——非平行超平面支持向量机得到了发展，并引起了人们的广泛兴趣。为解决繁重的训练时间成本限制了 SVM 处理大规模数据集的能力以及对于处理 XOR 数据集时，单个分离超平面的 SVM 预测性能往往不理想的问题，Jayadeva、Khemchandani 和 Chandra [2]构建了一对非平行超平面来区分新样本，称为孪生支持向量机(Twin support vector machine, TWSVM)，TWSVM 使得训练样本同时靠近自己和远离对方的超平面。由于 TWSVM 只要求解两个较小的 QPP，因此它的运行速度大约是标准支持向量机的四倍。由于 TWSVM 的良好性能，许多文献都致力于改进或发展它。Kumar 和 Gopal [3]用等式约束代替不等式约束，提出了最小二乘 TWSVM (LSTWSVM)，更加快了 TWSVM 的计算速度。但是 TWSVM 在训练和预测过程中存在不一致的问题，因此 Shao [4]等人设计了一种新的非平行超平面 SVM (NHSVM)，该支持向量机通过单个支持向量机优化问题可以获得两个分离超平面，使得支持向量机在训练和预测中是一致的。Qi Kai 等人[5]指出了 NHSVM 在几何视图上没有很

好的定义并提出了一种新的最小二乘几何非平行超平面支持向量机(LS-GNHSVM),但是 LS-GNHSVM使用的是凸损失函数,会导致模型的性能受异常值的影响而降低。Liu [6]提出了一种解决异常值和离群点敏感问题的 SVM, Ramp Loss LSSVM (RLSSVM)。

本文在 LS-GNHSVM 和 Ramp LossLSSVM 的启发下构造了一个基于 Ramp Loss 的 LS-GNHSVM,叫做 RLS-GNHSVM,应用于对信用卡客户违约概率的预测分析中,提高预测精度。其主要思想是,将 LS-GNHSVM 中的损失函数替换为斜坡损失,旨在解决 LS-GNHSVM 中凸损失函数对离群点和异常值的敏感问题。由于所提的 RLS-GNHSVM 是一个非凸优化问题,用 CCCP 算法对其转换成类似凸优化问题进行求解,用 CVXR 包求解模型。改进的斜坡损失最小二乘几何非平行超平面支持向量机(RLS-GNHSVM)对于异常值具有鲁棒性,为金融机构对客户的信用风险评估提供了更有力的工具。

2. 相关原理

在本章中,依次介绍了斜坡损失 SVM、斜坡损失 LSSVM 和 LS-GNHSVM。

2.1. 斜坡损失 SVM

对于有 n 个训练样本并且维度是 p 维的二分类问题,将第 i 个数据点定义为 $(x_i^T, y_i)^T$, 其中 $x_i \in \mathbb{R}^p$ 是数据输入值, $y_i \in \{+1, -1\}$ 是 x_i 对应的类别标签。正类数据有 n_+ 个, 负类数据有 n_- 个, 且 $n = n_+ + n_-$ 。定义 $X_+ \in \mathbb{R}^{n_+ \times p}$ 为正类数据集矩阵, $X_- \in \mathbb{R}^{n_- \times p}$ 为负类数据集矩阵, $Y = (y_1, \dots, y_n)^T$ 为标签向量。不做声明时,所有的向量为列向量。

标准支持向量机的目标是找到一个单独的超平面来适当地分离两类训练样本。最优分类器主要根据两个边界超平面之间的几何间隔最大的原则来确定。换句话说,必须优化以下问题:

$$\min_{w, b, \xi} \frac{1}{2} w^T w + c e^T \xi \quad (1)$$

其中, $X \in \mathbb{R}^{n \times p}$ 是数据矩阵, $D = \text{diag}(y_1, \dots, y_n)$ 是由数据类别标签构成的对角矩阵, 并且 e 是全为 1 的 n 维向量。 $w \in \mathbb{R}^p$ 和 $b \in \mathbb{R}$ 分别是决策超平面的法向量和截距项。 $c > 0$ 是调整参数, $\xi = (\xi_1, \dots, \xi_n)^T$ 是松弛变量, 引入松弛变量是为了允许训练样本违反原始不等式约束。

标准的 SVM 依赖于经典 Hinge 损失函数(见图 1(b))

$$H_s(z) = \max(0, s - z),$$

其中, 下标 s 表示 Hinge 点的位置, 以惩罚分类错误的样本, 则标准 SVM 的原问题可以进一步写成:

$$\min_{w, b, \xi} \frac{1}{2} w^T w + c e^T H_1(D(Xw + eb)) \quad (2)$$

由于 Hinge 损失函数的引入, 标准 SVM 对于异常观测值具有敏感性。为了提高 SVM 的鲁棒性, 避免将异常值转化为支持向量, 提出了斜坡损失函数[7](见图 1(a)), 也称为鲁棒的 Hinge 损失函数。

$$R_s(z) = \begin{cases} 0, & z > 1 \\ 1 - z, & s \leq z \leq 1 \\ 1 - s, & z < s \end{cases} \quad (3)$$

通过使 z 的值小于预定值 $s < 1$ 时的损失函数平坦化来取代 Hinge 损失函数。 $R_s(z)$ 可以分解为凸 Hinge 损失和凹损失(见图 1(c))之和,

$$R_s(z) = H_1(z) - H_s(z),$$

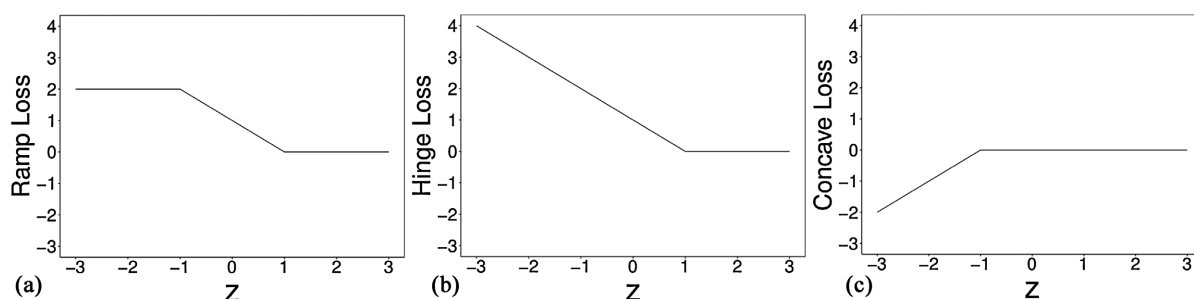


Figure 1. (a) Ramp loss; (b) Hinge loss; (c) Concave loss

图 1. (a) 斜坡损失函数; (b) Hinge 损失函数; (c) 凹损失函数

因此, 斜坡损失 SVM (RSVM) 的原问题可以写为:

$$\begin{aligned} \min_{w,b} & \frac{1}{2} w^T w + c e^T R_s(D(Xw + eb)) \\ &= \underbrace{\frac{1}{2} w^T w + c e^T H_1(D(Xw + eb))}_{\text{凸函数}} - \underbrace{c e^T H_s(D(Xw + eb))}_{\text{凹函数}} \end{aligned} \quad (4)$$

这可以通过 CCCP 过程[8]来求解。

2.2. 斜坡损失 LSSVM

尽管标准的 SVM (1) 在多种情况下都表现良好, 但是他的计算复杂度是 $O(n^3)$, 其中 n 是训练样本的数量。因此, 标准的 SVM 受限于解决大样本的数据。为解决这一问题, 提出了最小二乘支持向量机 LSSVM, LSSVM 将 SVM 的不等式约束代替为等式约束, LSSVM 考虑的是如下优化问题:

$$\begin{aligned} \min_{w,b,\xi} & \frac{1}{2} w^T w + c \xi^T \xi \\ \text{s.t.} & D(Xw + eb) = e - \xi, \end{aligned} \quad (5)$$

由于位于两个超平面附近的点的支持值最小, 它们对决策函数的贡献较小[8], 同时, 对于位于原理两个超平面的点, 特别是离群点, 它们往往有较大的支持值, 而我们想要消除这类点的影响。因此, RLSSVM 参照 ε -支持向量回归中加入 ε -不敏感损失函数的思想, 构造了如下优化问题:

$$\min_{w,b} \frac{1}{2} w^T w + \frac{c}{2} R_{\varepsilon,t}(e - D(Xw + eb)),$$

其中, $R_{\varepsilon,t}(z)$ 是 ε -不敏感斜坡损失函数(见图 2(a)),

$$R_{\varepsilon,t}(z) = \begin{cases} (t - \varepsilon)^2, & |z| > t \\ (|z| - \varepsilon)^2, & \varepsilon \leq |z| \leq t \\ 0, & |z| < \varepsilon \end{cases} \quad (6)$$

使得 ε -不敏感的二次损失函数(见图 2(b))

$$I_{\varepsilon}(z) = \max\{0, |z| - \varepsilon\}^2$$

在 $|z| > t > \varepsilon$ 时是光滑的。显然 ε -不敏感斜坡损失函数 $R_{\varepsilon,t}(z)$ 可以分解为凸 ε -不敏感二次损失函数和凹损失函数(见图 2(c))的和,

$$R_{\varepsilon,t}(z) = I_{\varepsilon}(z) - I_t(z). \quad (7)$$

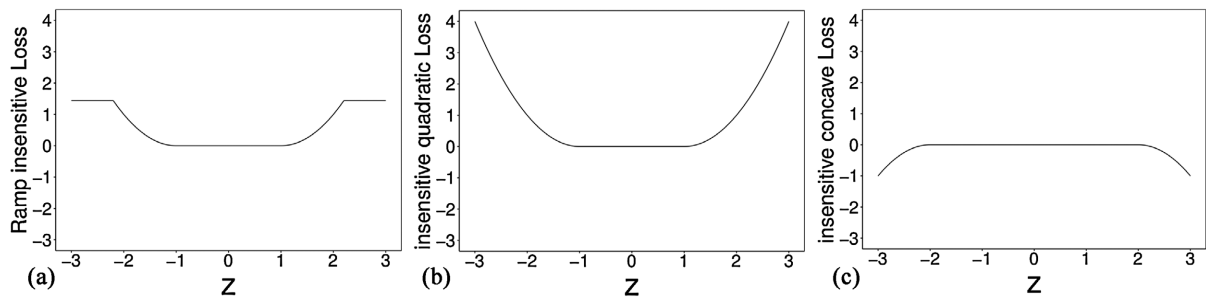


Figure 2. (a) Ramp ε -insensitive Loss; (b) ε -insensitive quadratic loss; (c) ε -insensitive concave loss

图 2. (a) ε -不敏感斜坡损失函数; (b) 凸 ε -不敏感二次损失函数; (c) 凹损失函数

2.3. LS-GNHSVM

另一种减少计算负担的方法是训练两个较小规模的样本,即所谓的孪生支持向量机(Twin Support Vector Machine, TWSVM)。与标准 SVM 和 LSSVM 不同的是, TWSVM 求解的是对应于每个类的一对非平行超平面。逻辑上,一个合理的分类器应该在训练和预测过程中被一致地定义。然而, TWSVM 在训练过程中没有考虑到两个分离超平面之间的距离。这个问题可能会降低性能,特别是在异方差的情况下。为了解决这个问题, Shao 等[4]提出了一种新的非平行超平面支持向量机(NHSVM)。尽管 NHSVM 是一致的,但是 Qi Kai 等人[5]发现模型 NHSVM 在几何视图中没有很好的定义。因此提出了最小二乘版本下的 GNHSVM。对于任意一对超平面,总是可以得到一个平均超平面:

$$\bar{l} : (w_+ + w_-)^T x + (b_+ + b_-) = 0.$$

其中, w_+ 和 b_+ 是正超平面和负超平面的法向量和截距项。如果样本 x_i 越靠近正(负)类超平面,则表明它等价的位于平均超平面的上(下)方。因此,样本到每个超平面的距离比较可以转化为相对于平均超平面的位置比较。则与 NHSVM 不同, LS-GNHSVM 在几何上更合理的约束可以写为:

$$\begin{cases} X_+ (w_+ + w_-) + e_+ (b_+ + b_-) \geq e_+ - \xi_+, \\ X_- (w_+ + w_-) + e_- (b_+ + b_-) \leq \xi_- - e_-, \\ \xi_+ \geq 0, \xi_- \geq 0, \end{cases} \quad (8)$$

这时约束条件(8)有两个优势:一是该约束与预测过程是一致的。根据约束(8),正(负)样本被限制在平均超平面的上方(下方),或者说等价地更靠近正(负)超平面。因此,约束条件(8)与预测过程的思想完全一致。二是该约束与最小化 $(\|X_+ w_+ + e_+ b_+\|^2 + \|X_- w_- + e_- b_-\|^2)$ 是兼容的。则 LS-GNHSVM 写为:

$$\begin{aligned} \min_{w_+, b_+, \xi_+, \xi_-} & \frac{C_1}{2} (\|X_+ w_+ + e_+ b_+\|^2 + \|X_- w_- + e_- b_-\|^2) \\ & + \frac{1}{2} (\|w_+\|^2 + b_+^2 + \|w_-\|^2 + b_-^2) + \frac{C_2}{2} (\|\xi_+\|^2 + \|\xi_-\|^2) \\ \text{s.t.} & \begin{cases} X_+ w_+ + e_+ b_+ + X_+ w_- + e_+ b_- = e_+ - \xi_+, \\ X_- w_- + e_- b_- + X_- w_+ + e_- b_+ = \xi_- - e_-. \end{cases} \end{aligned} \quad (9)$$

3. 斜坡损失最小二乘几何非平行超平面支持向量机

本文构造了结合 ε -不敏感斜坡损失函数下的在几视图上有较好的定义的非平行超平面支持向量机,称为斜坡损失最小二乘非平行超平面支持向量机(Ramp Loss LS-GNHSVM, RLS-GNHSVM)。在几何视图上有很好的定义并且能够解决二次损失函数导致的对于异常值敏感的问题以及使得支持向量机更具有稀

疏性。

3.1. 线性情况

$Q(z)$ 是二次损失函数, $Q(z) = z^T z$ 。为消除离群点对支持向量机的影响, 则我们用 ε -不敏感斜坡损失函数 $R_{\varepsilon,t}(z)$ 代替 $Q(z)$ 构造了如下优化问题:

$$\begin{aligned} \min_{\tilde{w}_{\pm}} & \underbrace{\frac{c_1}{2} \left((\tilde{X}_+ \tilde{w}_+)^T \tilde{X}_+ \tilde{w}_+ + (\tilde{X}_- \tilde{w}_-)^T \tilde{X}_- \tilde{w}_- \right) + \frac{1}{2} (\tilde{w}_+^T \tilde{w}_+ + \tilde{w}_-^T \tilde{w}_-)}_{\text{凸函数}} \\ & + \underbrace{\frac{c_2}{2} \left[I_{\varepsilon} (e_+ - D_+ f^+(X_+)) + I_{\varepsilon} (e_- - D_- f^-(X_-)) \right]}_{\text{凸函数}} \\ & - \underbrace{\frac{c_2}{2} \left[I_t (e_+ - D_+ f^+(X_+)) + I_t (e_- - D_- f^-(X_-)) \right]}_{\text{凹函数}}. \end{aligned} \quad (10)$$

其中, $\tilde{X}_+ = (X_+, e_+)$, $\tilde{X}_- = (X_-, e_-)$, $\tilde{w}_+^T = (w_+^T, b_+^T)$, $\tilde{w}_-^T = (w_-^T, b_-^T)$,

$$D = \begin{pmatrix} D_+ & \mathbf{0} \\ \mathbf{0} & D_- \end{pmatrix}, f^i(X) = \tilde{X}_i (\tilde{w}_+ + \tilde{w}_-), i = +, -, f(X) = \begin{pmatrix} f^+(X) \\ f^-(X) \end{pmatrix}.$$

根据 CCCP 框架[6]以及 LS-GNHSVM 则可以得出问题(10)的对偶问题。

其中, $\varepsilon \geq 0$ 是先验参数, $\eta_+ \in \mathbb{R}^{n_+}$, $\eta_- \in \mathbb{R}^{n_-}$, $\eta_+^* \in \mathbb{R}^{n_+}$, $\eta_-^* \in \mathbb{R}^{n_-}$ 。为了简化起见, 我们引入了符号:

$$\begin{aligned} \theta_i &= -\frac{c_2}{2} y_i \frac{\partial I_t(y_i f(x_i) - 1)}{\partial y_i f(x_i) - 1} \\ &= \begin{cases} -c_2 (|f(x_i) - y_i| - t), & y_i f(x_i) - 1 > t \\ c_2 (|f(x_i) - y_i| - t), & y_i f(x_i) - 1 < -t \\ 0, & \text{其他} \end{cases} \end{aligned}$$

其中 $i = 1, \dots, n$, $\theta = (\theta_+^T \quad \theta_-^T)^T$ 是由 θ_i 构成的列向量。则对偶问题可写为:

$$\begin{aligned} \min_{\tilde{\gamma}, \tilde{\gamma}} & \frac{1}{2} (\bar{\gamma} - \tilde{\gamma})^T Q^* (\bar{\gamma} - \tilde{\gamma}) + \frac{\theta^T (\bar{\gamma} - \tilde{\gamma})}{c_2} - D e^T (\bar{\gamma} - \tilde{\gamma}) + \varepsilon (\bar{\gamma} + \tilde{\gamma})^T e \\ \text{s.t.} & \begin{cases} (\bar{\gamma} - \tilde{\gamma})^T e = 0, \\ \bar{\gamma} \geq -\bar{\theta}, \\ \tilde{\gamma} \geq \tilde{\theta}. \end{cases} \end{aligned} \quad (11)$$

其中, $\gamma_+ \in \mathbb{R}^{n_+}$, $\gamma_- \in \mathbb{R}^{n_-}$, $\gamma_+^* \in \mathbb{R}^{n_+}$, $\gamma_-^* \in \mathbb{R}^{n_-}$ 是拉格朗日乘子向量, I_{p+1} 是 $p+1$ 阶单位阵。

$$\tilde{X} = \begin{pmatrix} \tilde{X}_+ \\ \tilde{X}_- \end{pmatrix}, \eta = \begin{pmatrix} \eta_+ \\ \eta_- \end{pmatrix}, \eta^* = \begin{pmatrix} \eta_+^* \\ \eta_-^* \end{pmatrix}, \gamma = \begin{pmatrix} \gamma_+ \\ \gamma_- \end{pmatrix}, \gamma^* = \begin{pmatrix} \gamma_+^* \\ \gamma_-^* \end{pmatrix}.$$

$$\theta_i = \bar{\theta}_i + \tilde{\theta}_i, \quad i = 1, \dots, n. \quad Q = \tilde{X} \left[(I_{p+1} + c_1 \tilde{X}_+^T \tilde{X}_+)^{-1} + (I_{p+1} + c_1 \tilde{X}_-^T \tilde{X}_-)^{-1} \right] \tilde{X}^T, \quad Q^* = Q + \frac{I_n}{c_2}, \quad I_n \text{ 是 } n \text{ 阶单位阵},$$

$\bar{\gamma}_i = \gamma_i^* - \bar{\theta}_i$, $\tilde{\gamma}_i = \gamma_i + \tilde{\theta}_i$, $i = 1, 2, \dots, n$ 。

综上所述, 给出了本文所提方法的 CCCP 求解过程如表 1 所示:

Table 1. The process of solving model parameters**表 1.** 求解模型参数过程**RLS-GNHSVM**

- (1) 输入训练样本 $X_+ \in \mathbb{R}^{n_+ \times p}$, $X_- \in \mathbb{R}^{n_- \times p}$ 及其对应标签 $Y = (y_1, \dots, y_n)^T$;
- (2) 选择参数 $\varepsilon, t > 0$, 选择适当的惩罚参数 $c_1, c_2 > 0$ 和核函数 $K(x, x^T)$, 初始化 $\bar{\theta}^0, \tilde{\theta}^0$, 设 $k=1$;
- (3) 构造并求解第 k 个迭代步骤中的 QPP 问题:

$$\min_{\tilde{\gamma}, \tilde{\gamma}} \frac{1}{2} (\bar{\gamma} - \tilde{\gamma})^T Q^* (\bar{\gamma} - \tilde{\gamma}) + \frac{\theta^T (\bar{\gamma} - \tilde{\gamma})}{c_2} - D e^T (\bar{\gamma} - \tilde{\gamma}) + \varepsilon (\bar{\gamma} + \tilde{\gamma})^T e$$

$$\text{s.t.} \begin{cases} (\bar{\gamma} - \tilde{\gamma})^T e = 0, \\ \bar{\gamma} \geq -\bar{\theta}, \\ \tilde{\gamma} \geq \tilde{\theta}. \end{cases}$$

得到 $(\tilde{\gamma}^k, \bar{\gamma}^k)$, 计算 $(\tilde{w}_+^k, \tilde{w}_-^k)$,

则决策函数为 $f^+(X_+) = X_+ (w_+ + w_-) + (b_+ + b_-)$, $f^-(X_-) = X_- (w_+ + w_-) + (b_+ + b_-)$;

- (4) 计算 $(\bar{\theta}^k, \tilde{\theta}^k)$;
- (5) 若 $(\bar{\theta}^k, \tilde{\theta}^k) = (\bar{\theta}^{k-1}, \tilde{\theta}^{k-1})$, 则输出(3)中得到的 w_+^k, b_+^k , 运行第(6)步, 否则令 $k = k+1$, 返回第(3)步;
- (6) 给一个新的样本点 x_{new} , 应用如下决策函数为其进行分类:

$$\text{Class } i = \min_{k=+, -} \frac{|w_k^T x_{\text{new}} + b_k|}{\|w_k\|}.$$

3.2. 非线性情况

接下来, 我们将 RLS-GNHSVM 从线性情况推广到非线性的场景下, 设 $\phi(\cdot)$ 是原始样本空间到某个高维希尔伯特空间的映射, $K(\cdot, \cdot)$ 是其对应的核函数。因此, 对于任意两个样本点 x_1, x_2 , 都有 $K(x_1^T, x_2^T) = \phi(x_1)^T \phi(x_2)$ 。

在经过核变换后, 我们考虑如下两个非平行分离超平面:

$$K(x^T, X^T) w_+ + b_+ = 0, \quad K(x^T, X^T) w_- + b_- = 0,$$

$$\text{令 } \tilde{X}_{K+} = (K(X_+, X^T), e_+), \tilde{X}_{K-} = (K(X_-, X^T), e_-), \tilde{w}_+ = (w_+^T, b_+)^T, \tilde{w}_- = (w_-^T, b_-)^T,$$

则非线性的 RLS-GNHSVM 定义为:

$$\begin{aligned} \min_{\tilde{w}_\pm, \eta_\pm^*} & \frac{1}{2} (\tilde{w}_+^T \tilde{w}_+ + \tilde{w}_-^T \tilde{w}_-) + \frac{c_1}{2} \left((\tilde{X}_{K+} \tilde{w}_+)^T \tilde{X}_{K+} \tilde{w}_+ + (\tilde{X}_{K-} \tilde{w}_-)^T \tilde{X}_{K-} \tilde{w}_- \right) \\ & + \frac{c_2}{2} \left(\|\eta_+\|^2 + \|\eta_-\|^2 + \|\eta_+^*\|^2 + \|\eta_-^*\|^2 \right) + \theta_+^T (\tilde{X}_{K+} \tilde{w}_+ + \tilde{X}_{K+} \tilde{w}_-) \\ & + \theta_-^T (\tilde{X}_{K-} \tilde{w}_- + \tilde{X}_{K-} \tilde{w}_-) \\ \text{s.t.} & \begin{cases} \tilde{X}_{K+} \tilde{w}_+ + \tilde{X}_{K+} \tilde{w}_- = \varepsilon e_+ + \eta_+ + e_+, \\ \tilde{X}_{K-} \tilde{w}_- + \tilde{X}_{K-} \tilde{w}_+ = \varepsilon e_- + \eta_- - e_-, \\ \tilde{X}_{K+} \tilde{w}_+ + \tilde{X}_{K+} \tilde{w}_- = e_+ - \varepsilon e_+ - \eta_+^*, \\ \tilde{X}_{K-} \tilde{w}_- + \tilde{X}_{K-} \tilde{w}_+ = -\varepsilon e_- - e_- - \eta_-^*. \end{cases} \end{aligned} \quad (12)$$

其对偶问题为:

$$\begin{aligned} \min_{\bar{\gamma}, \tilde{\gamma}} \quad & \frac{1}{2} (\bar{\gamma} - \tilde{\gamma})^T Q_K^* (\bar{\gamma} - \tilde{\gamma}) + \frac{\theta^T (\bar{\gamma} - \tilde{\gamma})}{c_2} - D e^T (\bar{\gamma} - \tilde{\gamma}) + \varepsilon (\bar{\gamma} + \tilde{\gamma})^T e \\ \text{s.t.} \quad & \begin{cases} (\bar{\gamma} - \tilde{\gamma})^T e = 0, \\ \bar{\gamma} \geq -\bar{\theta}, \\ \tilde{\gamma} \geq \tilde{\theta}. \end{cases} \end{aligned} \quad (13)$$

其中, $Q_K^* = \tilde{X}_K \left[(I_{p+1} + c_1 \tilde{X}_{K+}^T \tilde{X}_{K+})^{-1} + (I_{p+1} + c_1 \tilde{X}_{K-}^T \tilde{X}_{K-})^{-1} \right] \tilde{X}_K^T + \frac{I_n}{c_2}$,

$$\tilde{X}_K = \begin{pmatrix} \tilde{X}_{K+} \\ \tilde{X}_{K-} \end{pmatrix}, \bar{\gamma} = \begin{pmatrix} \bar{\gamma}_+ \\ \bar{\gamma}_- \end{pmatrix}, \tilde{\gamma} = \begin{pmatrix} \tilde{\gamma}_+ \\ \tilde{\gamma}_- \end{pmatrix}, e = \begin{pmatrix} e_+ \\ e_- \end{pmatrix}, \theta = \begin{pmatrix} \theta_+ \\ \theta_- \end{pmatrix}, \theta = \tilde{\theta} + \bar{\theta}$$

我们可以发现, 优化对偶问题(13)与线性场景下求解问题(11)是类似的。类似地, 得到 $w_{\pm}, b_{\pm}$ 后, 可以通过如下决策函数对新加入的样本点进行分类,

$$\text{Class } i = \min_{k=+,-} \frac{|K(x_{\text{new}}^T, X^T) w_k + b_k|}{\sqrt{w_k^T K(X, X^T) w_k}}.$$

4. 基于 RLS-GNHSVM 的信用卡客户违约概率预测模型

4.1. 数据来源与指标设计

本文的数据来自 UCI 机器学习库(<https://archive.ics.uci.edu/>)公开的数据, 该数据集台湾地区的信用卡客户拖欠付款情况的数据, 通常, 从风险管理的角度来看, 预测违约概率的预测准确性的结果将比分类的二元结果(可信或不可信的客户)更有价值。由于违约的真实概率未知, 本研究提出了一种新的稳健的分类方法来估计违约的真实概率。每条数据样本均有 23 个属性。本文的响应变量为客户是否付款违约, 一共有两种可能的状态: 违约和不违约。综合考虑数据结构和指标特征, 在众多影响因素中主要考虑了以下 23 个解释变量对响应变量进行建模分析, 这些变量均与客户是否存在付款违约有高度相关关系[9]。变量具体说明如表 2。

(1) 年龄(age): 该指标是信用卡客户的年龄, 用于反映客户所处年龄阶段。它是一个多样化的指标, 涵盖了从青年到老年的广泛年龄段。不同年龄段的客户对信用卡的需求和使用习惯有所不同, 但总体上, 信用卡作为一种支付工具, 在各年龄段中均有一定的用户基础。年轻客户可能更看重信用卡的消费便利性和优惠活动, 而中年和老年客户则可能更注重信用卡的支付安全性、信用积累以及附加服务。因此, 年龄是理解和分析信用卡客户需求的一个重要维度。

(2) 婚姻(marital): 该指标是客户的婚姻状况, 有助于银行更全面地了解客户的个人情况和生活状态。婚姻状况在一定程度上反映了客户的经济稳定性和责任感。已婚人士可能因家庭责任而更加注重财务管理和信用记录, 这可能会对信用评分产生积极影响。针对不同婚姻状况的客户, 银行可能会制定差异化的营销策略。例如, 对于已婚已育的客户, 银行可能会推出更多与家庭消费相关的优惠活动; 而对于未婚或离婚的客户, 则可能更注重提供个性化的信用服务和产品。

(3) 性别(gender): 该指标是客户的性别, 信用卡客户的性别是银行在数据分析和市场细分时考虑的一个基础指标。不同性别的客户在信用卡消费习惯上可能存在一定的差异。性别也可能影响客户的风险偏好。虽然这并非绝对, 但一些研究指出, 女性客户在财务决策上可能更为谨慎, 更注重长期规划和风险控制; 而男性客户则可能更倾向于追求高风险高收益的投资机会。这种差异对于银行在提供信用卡

Table 2. Data indicator variables
表 2. 数据指标变量

变量类型	变量名	详细说明	取值范围
因变量	是否违约罚款	客户信用卡还款拖欠情况	0: 不违约 1: 违约
	年龄(age)	客户年龄, 反映客户所处年龄阶段。	20~80
	婚姻(marital)	婚姻状况: “已婚”, “单身”, “其他”	1: 已婚 2: 单身 3: 其他
	性别(gender)	性别: “男”, “女”	1: 男 2: 女
	受教育程度(education)	受教育程度: “研究生”, “大学”, “高中”, “其他”	1: 研究生 2: 大学 3: 高中 4: 其他
解释变量	给予信贷金额 (Amount of the given credit)	个人消费信贷和家人信贷	取值为正整数
	过去付款情况 (History of past payment)	过去 5 个月(2005 年 4~9 月)的每月还款情况	-1: “按时支付” 1: “延迟付款 1 个月” 2: “延迟付款 2 个月” ... 9: “延迟付款 9 个月”
	账单金额 (Amount of bill statement)	过去 5 个月(2005 年 4~9 月)的每月账单金额	取值为整数
	过去付款金额 (Amount of previous payment)	过去 5 个月(2005 年 4~9 月)的每月付款金额	取值是整数

服务时, 如何平衡风险与收益具有参考价值。

(4) 受教育程度(education): 该指标是信用卡客户的受教育程度, 它是评估其信用风险、消费习惯及还款能力的重要指标之一。受教育程度较高的客户通常具备更强的财务管理能力和风险意识, 他们更有可能清晰理解信用卡的使用规则、利息计算方式及逾期费用等, 从而避免不必要的支出和逾期情况。在实际应用中, 银行会根据客户的具体情况进行审批和额度授予。例如, 对于受教育程度较低但收入稳定、信用记录良好的客户, 银行也可能会给予较高的信用额度。

(5) 给予信贷金额(Amount of the given credit): 该指标包括个人消费信贷和个人信贷。在信用卡的语境中, 信贷金额通常指的是银行或金融机构为信用卡客户设定的信用额度, 即客户可以在一定期限内透支或使用的最大金额。这个额度是银行根据客户的信用状况、收入水平、还款能力等因素综合评估后确定的。

(6) 过去付款情况(History of past payment): 数据跟踪了过去(2005 年 4 月至 9 月)的每月还款记录如下: 2005 年 9 月的还款情况; 2005 年 8 月的还款情况; ...; 2005 年 4 月的还款状态。还款状况的衡量标准为: -1 = 按时支付; 1 = 延迟付款一个月; 2 = 延迟付款两个月; ...; 8 = 延迟付款八个月; 9 = 延迟付款 9 个月及以上。

(7) 账单金额(Amount of bill state-ment): 数据跟踪了过去(2005 年 4 月至 9 月)每月客户的信用卡账单金额。

(8) 过去付款金额(Amount of previous payment): 数据跟踪了过去(2005 年 4 月至 9 月)每月客户支付的信用卡金额。

4.2. 信用卡还款违约概率预测模型的建立

在本节中进行实证研究, 将提出的 RLS-GNHSVM 应用于信用卡还款违约概率预测模型, 实验都在 R4.4.1 中进行, 在基于系统配置 Intel(R) Core(TM) i7-10700 CPU(2.90 GHz)和 16 GB RAM 的 PC 上运行, 使用优化工具包 CVXR 求解模型。关于模型参数的设置, 对于 LSTSVM、NHSVM、LS-GNHSVM 和 RLS-GNHSVM, 设置 $C = \{2^{-8}, 2^{-7}, \dots, 2^7, 2^8\}$, $C_1 = \{2^{-8}, 2^{-7}, \dots, 2^7, 2^8\}$, $C_2 = \{2^{-8}, 2^{-7}, \dots, 2^7, 2^8\}$; 设置 RLS-GNHSVM 中 ε, t 的取值范围为 $\varepsilon = \{0.1, 0.2, 0.3, 0.4, 0.5\}$, $t = \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ 。非线性情况, 考虑高斯核函数, 评价指标为准确率(Acc)。对于所有模型参数的选择, 本文使用五折交叉验证法来搜索模型的最优参数。为了消除实验随机性的影响, 进行 100 次数值实验, 计算评价指标的平均值(Acc)和标准差(sd)。考虑实际数据为非线性可分, 所以建立高斯核函数下非线性可分模型, 模型结果如表 3 所示。提出的模型提高了预测准确率, 将其应用于客户信用卡还款违约概率预测模型可以带来更好的效果。

Table 3. Data set prediction results

表 3. 数据集预测结果

	LSTSVM Acc. $\pm$ sd	NHSVM Acc. $\pm$ sd	LS-GNHSVM Acc. $\pm$ sd	RLS-GNHSVM Acc. $\pm$ sd
数据集	0.510 $\pm$ 0.029	0.825 $\pm$ 0.015	0.829 $\pm$ 0.015	0.835 $\pm$ 0.015

5. 结论

本文聚焦于提升金融服务领域对信用卡违约概率评估的准确性和效率, 通过对 UCI 机器学习库中信用卡客户违约数据的深入分析, 构建并验证了一种创新的预测模型——改进的斜坡损失最小二乘几何非平行超平面支持向量机(RLS-GNHSVM)。该模型通过融合斜坡损失函数与最小二乘几何非平行超平面支持向量机的各自优势, 成功克服了传统凸损失函数在处理含噪数据或存在异常值时易受影响、性能下降的局限。在实证研究过程中, RLS-GNHSVM 模型展现了其卓越的稳定性和预测精度。相较于其他三种模型, RLS-GNHSVM 不仅在处理复杂多变的信用卡客户数据时表现出更高的鲁棒性, 而且在预测客户违约概率方面实现了显著的效能提升, 为金融机构提供了一种更为精准、高效的风险评估工具。

参考文献

- [1] Vapnik, V.N. (1999) An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, **10**, 988-999. <https://doi.org/10.1109/72.788640>
- [2] Jayadeva, Khemchandani, R. and Chandra, S. (2007) Twin Support Vector Machines for Pattern Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 905-910. <https://doi.org/10.1109/tpami.2007.1068>
- [3] Arun Kumar, M. and Gopal, M. (2009) Least Squares Twin Support Vector Machines for Pattern Classification. *Expert Systems with Applications*, **36**, 7535-7543. <https://doi.org/10.1016/j.eswa.2008.09.066>
- [4] Shao, Y., Chen, W. and Deng, N. (2014) Nonparallel Hyperplane Support Vector Machine for Binary Classification Problems. *Information Sciences*, **263**, 22-35. <https://doi.org/10.1016/j.ins.2013.11.003>
- [5] Qi, K. and Yang, H. (2023) LS-GNHSVM: A Novel Joint Geometrical Nonparallel Hyperplane Support Vector Machine. *Expert Systems with Applications*, **215**, 119413. <https://doi.org/10.1016/j.eswa.2022.119413>
- [6] Liu, D., Shi, Y., Tian, Y. and Huang, X. (2016) Ramp Loss Least Squares Support Vector Machine. *Journal of Computational Science*, **14**, 61-68. <https://doi.org/10.1016/j.jocs.2016.02.001>
- [7] Collobert, R., Sinz, F., Weston, J. and Bottou, L. (2006). Trading Convexity for Scalability. *Proceedings of the 23rd International Conference on Machine Learning - ICML'06*, 201-208.

- <https://doi.org/10.1145/1143844.1143870>.
- [8] Yuille, A.L. and Rangarajan, A. (2003) The Concave-Convex Procedure. *Neural Computation*, **15**, 915-936.
<https://doi.org/10.1162/08997660360581958>
- [9] Yeh, I. and Lien, C. (2009) The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Systems with Applications*, **36**, 2473-2480.
<https://doi.org/10.1016/j.eswa.2007.12.020>