

电商背景下的在线购买者购买意向预测分析

张 朋

贵州大学数学与统计学院, 贵州 贵阳

收稿日期: 2025年8月14日; 录用日期: 2025年8月25日; 发布日期: 2025年9月24日

摘 要

在人工智能与数据科学快速发展的背景下, 众多机器学习方法相继涌现, 其中支持向量机凭借其坚实的理论基础成为广泛应用的分类模型。然而, 在实际应用中, 当数据集存在噪声或异常值时, 传统支持向量机的分类性能下降。本研究基于UCI机器学习库的在线购物者购买意向数据集(Online Shoppers Purchasing Intention Dataset), 该数据集涵盖12,330个不同用户的会话记录, 时间跨度为一年, 包含17个多维度特征, 涉及页面访问次数与停留时长、跳出率、退出率、页面价值、节日接近度, 以及访问来源、访客类型、访问月份等类别变量。目标变量为是否产生购买意向(二分类)。本文创新性地提出了一种基于有界框架的有界弹球孪生支持向量机的鲁棒分类器模型。本研究以电商平台的在线购买者行为数据为基础, 针对电商背景下的购买意向预测问题, 提出并应用了基于有界分位损失的孪生支持向量机(BP-TSVM)模型。数据集涵盖用户在电商网站的多维访问特征及历史行为记录, 目标变量为是否产生购买意向(二分类)。为验证模型的有效性, 本文将BP-TSVM与经典孪生支持向量机(TSVM)、弹球孪生支持向量机(PinTSVM)以及最小二乘孪生支持向量机(TPMSVM)进行对比实验。实验在相同电商数据集和参数优化策略下进行, 并采用准确率、sd值等指标评估模型性能。结果表明, BP-TSVM在预测精度和鲁棒性方面均优于对比模型, 尤其在处理存在噪声的电商购买意向数据时表现更为稳定。该研究为电商企业在精准营销、用户分层及个性化推荐等方面提供了可行的技术路径和实证依据。

关键词

有界框架, 弹球损失, 鲁棒性, 有界弹球孪生支持向量机, 在线购物者购买意向, 电商

Predictive Analysis of Online Shoppers' Purchase Intention in the E-Commerce Context

Peng Zhang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Aug. 14th, 2025; accepted: Aug. 25th, 2025; published: Sep. 24th, 2025

文章引用: 张朋. 电商背景下的在线购买者购买意向预测分析[J]. 电子商务评论, 2025, 14(9): 1786-1796.

DOI: 10.12677/ecl.2025.1493104

Abstract

With the rapid advancement of artificial intelligence and data science, numerous machine learning techniques have emerged, among which Support Vector Machines (SVMs) have gained wide adoption for classification tasks due to their solid theoretical foundation. However, their performance often deteriorates in the presence of noise or outliers. This study utilizes the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository, comprising 12,330 unique user sessions collected over one year, with 17 behavioral and categorical features such as page visit counts and durations, bounce and exit rates, page values, proximity to special days, traffic sources, visitor types, and months of visit. The target is a binary variable indicating purchase intention. We propose a novel robust classifier, the Bounded Pinball Twin Support Vector Machine (BP-TSVM), designed for purchase intention prediction in e-commerce environments. The dataset includes multi-dimensional access characteristics and historical behavior records of users on e-commerce websites, with the target variable being whether there is a purchase intention (binary classification). The proposed method is benchmarked against the classical Twin Support Vector Machine (TSVM), Pinball Twin Support Vector Machine (PinTSVM), and Twin Parametric Margin Support Vector Machine (TPMSVM) under identical parameter optimization settings. Experimental evaluation using accuracy and standard deviation metrics shows that BP-TSVM consistently outperforms the baselines in both predictive accuracy and robustness, particularly on noisy e-commerce datasets. These findings offer a viable technical solution and empirical insights for enhancing precision marketing, customer segmentation, and personalized recommendation in e-commerce platforms.

Keywords

Bounded Framework, Pinball Loss, Bounded Pinball Twin Support Vector Machine (BP-TSVM), Online Shoppers' Purchase Intention, E-Commerce

Copyright © 2025 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. 背景

近年来,随着互联网技术和移动终端的快速发展,电子商务已成为全球零售业和服务业的重要组成部分。电商平台通过整合商品展示、支付结算、物流配送等功能,为消费者提供了便捷、高效的购物体验。与此同时,消费者在电商平台上的行为数据不断积累,为企业挖掘用户潜在需求、制定精准营销策略提供了丰富的数据资源。

在这一背景下,在线购买者的购买意向(Online Purchase Intention)成为电商领域的重要研究主题。购买意向是指消费者在特定情境下倾向于进行实际购买行为的心理状态,它不仅反映了用户的兴趣与需求,还直接影响转化率和销售额。影响购买意向的因素多样,既包括价格、促销、节日等外部环境变量,也涵盖网站可用性、产品信息质量、支付安全性等平台因素,还涉及消费者的个人特征与历史行为模式。

对于电商企业而言,准确预测在线购买者的购买意向具有重要意义。一方面,预测结果可用于提前识别潜在客户,从而优化营销资源配置,提升广告投放的精准度;另一方面,可以辅助平台在产品推荐、库存管理、个性化优惠等环节进行策略优化,提高用户体验并促进客户忠诚度。在大规模电商平台中,用户行为数据往往具有高维度、非线性、噪声干扰及类别不平衡等特点,这对预测模型的鲁棒性与泛化能力提出了更高要求。因此,研究和构建高精度、强鲁棒性的购买意向预测模型,已成为电商数据挖掘

与智能营销领域的核心任务之一。

支持向量机(Support Vector Machines, SVM)是由 Vapnik [1]等人(1995 年)基于统计学习理论中的 VC 理论和结构风险最小化理论提出的一种机器学习方法。支持向量机一经提出就被广泛应用于各个领域,如文本识别、人脸识别[2]-[4]。实际上,求解支持向量机就是求解一个带有约束问题的二次规划问题(Quadratic Programming Problem, QPP)。SVM 的训练算法一般采用的是将原来的大规模的二次规划问题分解成若干个小规模的二次规划问题,然后根据某种迭代方法对这些小规模的二次规划问题反复求解,最后根据得到的小规模二次规划问题的解来构造大规模二次规划问题的近似解[5]。然而这样的训练算法时间成本往往很高,因此,为了提高 SVM 的训练速度,研究者们提出了许多的改进算法。其中比较经典的改进算法有块算法(Chunking Algorithm) [6]、分解算法(Decomposition Algorithm) [7]和序列最小优化算法(Sequential Minimal Optimization, SMO) [8]等。

随着 SVM 理论的不完备和应用层面的扩展,其也逐渐暴露出一些局限性,例如对特征噪声和离群点敏感等问题。特征噪声通常在决策边界附近产生[9],因此特征噪声会对 SVM 中的支持向量的数量造成影响,从而影响决策超平面,导致相应的 SVM 分类器对噪声敏感,并且对于重采样不稳定。相比之下,弹球损失函数与分位数距离相关,对决策超平面附近的噪音不太敏感。Huang 等人将弹球损失函数应用于 SVM,提出了基于弹球损失函数的支持向量机(Pin-SVM) [10]。同时研究了 Pin-SVM 的性质,包括噪声不敏感性、稳健性和误分类误差最小化。与经典的 SVM 相比,Pin-SVM 具有相同的计算复杂度,并且具有噪声不敏感性和重采样稳定性。多位学者也提出了各种方法来解决 SVM 对特征噪声不稳健的问题。铰链损失的另一个缺点是对离群点(标签噪声)敏感,所以基于铰链损失的 SVM 获得的决策超平面容易被离群点影响。为了改进 SVM 的性能,Wu 和 Liu 构造了截断铰链损失函数,称为斜坡损失函数,同时证明了基于斜坡损失函数的 SVM 对离群点的稳健性[11]。在斜坡损失函数的基础上,Tang 等人提出了一种新的鲁棒性损失函数,并将其应用到 SVM 中进行多分类任务[12]。2007 年,源于 GEPSVM 的思想,Jayadeva [13]等人提出了孪生支持向量机(Twin Support Vector Machines, TWSVM)。TWSVM 将两个广义特征问题转化为两个较小的类似支持向量机的凸规划问题。相对于 SVM, TWSVM 将所有样本点都放在凸规划问题的约束条件中,而 TWSVM 将与目标函数有关的样本点的另一类样本点放在约束条件中。这样一来, TWSVM 的训练速度显著提高,其训练时间仅为 SVM 的 1/4。

随着互联网技术的深度普及与数字经济的蓬勃发展,电子商务已从一种新兴的商业模式演进为驱动国民经济高质量发展的重要引擎。对电商经济的研究不再局限于其交易效率的提升,更深入到其对宏观经济增长、微观商业模式创新以及市场结构变革的深刻影响。在此背景下,本研究旨在深入探讨电商经济的新特征、新动能及其演化机理。

从宏观层面看,电子商务通过降低交易成本、拓宽市场边界、促进消费升级,显著促进了经济增长、就业与产业融合。其作为一种深度融合互联网技术与金融服务的创新模式,为理解电商经济的价值创造奠定了理论基础[14]。

然而,电商经济的内涵远不止于宏观贡献。其生命力在于持续的模式创新与业态演化。近年来,以直播电商、社交电商为代表的新业态迅猛发展,彻底重塑了人、货、场的关系。新业态的成功并非偶然,其背后是网红与消费者之间新型信任关系与互动机制的建立,这极大地改变了传统的消费决策路径,构成了电商经济发展的新动能[15]。

这些不断涌现的新模式、新业态,本质上都依托于平台型企业的组织与创新。电商平台已从简单的交易中介演变为整合物流、支付、数据、社交的复杂生态圈。平台企业的商业模式创新,特别是其“属性-功能”认知的持续演化,是推动整个电商经济生态系统不断升级和扩张的核心驱动力[16]。

在此背景下,本研究将在前人工作的基础上,受 Pin-TSVM 和有界框架的启发,本文提出了一种新

的基于弹球损失的有界弹球孪生支持向量机(BP-TSVM)模型应用于购买者购买意向的预测模型,提高预测精度。主要思想是将有界框架与 Pin-TSVM 结合,将无界损失函数变成有界,减少噪声对决策超平面的影响,使 BP-TSVM 在处理噪声数据时具有稳健性。BP-TSVM 使用松弛变量的 L1 范数,可以使用拉格朗日乘子法求解对偶问题,然后通过 clip-DCD 算法求解模型。改进的基于弹球损失的有界弹球孪生支持向量机(BP-TSVM)模型对异常值具有鲁棒性。

2. 相关理论

假设 (x_i, y_i) , $i = 1, \dots, n$ 为训练样本, 其中 $D = \text{diag}(y_1, y_2, \dots, y_n)$

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T \in \mathbb{R}^p, \quad y_i \in \{-1, +1\}.$$

定义样本矩阵

$$X = (x_1 \quad x_2 \quad \dots \quad x_n)$$

并称其为样本矩阵。

当 $y_i = 1$ 时, 将 x_i 称为正类样本, 反之, 称为负类样本。

将正类样本与负类样本所构成的样本矩阵分别记为 X^+ 和 X^- , 其中

$$X^+ \in \mathbb{R}^{n^+ \times p}, \quad X^- \in \mathbb{R}^{n^- \times p},$$

并且 n^+ 和 n^- 分别代表正类样本数和负类样本数, 且有

$$n^- + n^+ = n.$$

2.1. 基于铰链损失的支持向量机

传统的 SVM 算法是基于铰链损失函数提出的, 铰链损失函数定义如下:

$$L_{\text{hinge}}(u) = \max\{0, u\}. \quad (1)$$

通过引入调节参数 $C (C > 0)$ 和松弛变量 $\xi_i, i = 1, 2, \dots, n$, SVM 可表示为如下形式

$$\min_{w, b, \xi} \frac{1}{2} w^T w + C e^T \xi, \quad \text{s.t.} \quad y^\circ (Xw + be) \geq e - \xi, \quad \xi \geq 0, \quad (2)$$

其中 $\xi = (\xi_1, \xi_2, \dots, \xi_n)^T$ 是松弛向量, C 是非负惩罚参数, e 和 0 分别是元素全为 1 和 0 的 n 维列向量。

根据拉格朗日乘子法及 KKT 条件, 可得(2)的对偶问题为:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T D X X^T D \alpha - e^T \alpha \\ \text{s.t.} \quad & \begin{cases} 0 \leq \alpha \leq C e, \\ e^T D \alpha = 0. \end{cases} \end{aligned} \quad (3)$$

其中 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$ 是拉格朗日乘子向量。若求得(3)的最优解 α , 即可确定超平面 $w^T x + b = 0$ 。

给定一个新的样本点 x_{new} , 其判别函数为:

$$f(x_{\text{new}}) = \text{sgn}(w^T x_{\text{new}} + b). \quad (4)$$

在非线性情况下, 定义映射 $\phi(\cdot)$:

$$x \xrightarrow{\phi} \phi(x),$$

并引入核函数：

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j), \quad (5)$$

通过核函数可将原本线性不可分的数据映射到高维空间，使数据线性可分，从而解决非线性分类问题。

2.2. 基于弹球损失的支持向量机

为了进一步提高支持向量机的性能，Huang *et al.* 使用分位数距离来度量裕度，并提出相应的分类器，通过弹球损失来最大化分位数距离，弹球损失函数定义如下：

其中 $0 < \tau \leq 1$ 是调节参数。引入弹球损失，Pin-SVM 表达式如下：

$$\min_{\omega, \xi} \frac{1}{2} \|\omega\|^2 + C e^T \xi \quad (6)$$

其中 ξ 是松弛向量， C 是非负的惩罚参数。同理 SVM，(6) 式对偶问题为

$$\begin{aligned} \min_u \quad & \frac{1}{2} u^T D X X^T D u - e^T u \\ \text{s.t.} \quad & \begin{cases} -C\tau e \leq u \leq Ce, \\ e^T D u = 0. \end{cases} \end{aligned} \quad (7)$$

Pin-SVM 的主要优点是对噪声不敏感，特别是对决策边界附近的特征噪声不敏感。

3. 基于有界弹球损失函数的支持向量机(BP-TSVM)

3.1. 基于有界弹球损失函数的支持向量机(BP-TSVM)

传统的分类方法在遇到噪声数据时效果变差。有界损失函数的应用是提高稳健性的有效手段。然而，大多数有界损失函数都以斜坡形式存在，由于硬截断，失去了原始函数的一些固有性质。此外，关于如何设计有界损失函数，目前还没有统一的框架。针对上述两个问题，Fu 等人提出了一个可以平滑自适应地约束任何非负函数的通用框架。它不仅可以退化为原来的函数，而且可以继承其良好的性质，包括对称性、可微性和光滑性。有界损失函数的框架如下：

$$L(x) = \frac{1}{\lambda} \left(1 - \frac{1}{1 + b h(x)} \right), \quad \forall \lambda, b > 0 \quad (8)$$

受有界损失函数框架启发，本文在有界损失函数的一般框架下构造了一种新的鲁棒损失函数，称为有界弹球损失(Bounded Pinball Loss, BP Loss)。

有界弹球损失函数如下：

$$L_{\text{BSP}}(u) = \frac{1}{\eta} \left(1 - \frac{1}{1 + \lambda u} \right) \text{ if } u \geq 0, \quad L_{\text{BSP}}(u) = \frac{1}{\eta} \left(1 - \frac{1}{1 + \lambda \tau u} \right) \text{ if } u < 0 \quad (9)$$

其中 $0 \leq \tau \leq 1$ ， $\eta > 0$ ， $\lambda > 0$ 是调节参数，且 τ 控制损失函数的不对称性。显然， $L_{\text{BP}}(u)$ 是有界的，这使得它对离群值不敏感。

3.2. 基于 BP 的 TSVM

Pin-TSVM 已被证明是一种有效的分类算法。然而，我们注意到它的弹球损失函数是无界的。为了提高泛化性能，我们基于有界弹球损失函数提供了一个稳定的 BP-TSVM，它也是通过求解一对二次

规划问题得到了一对非平行平面。

在线性情况下, BP-TSVM 在 $\mathbb{R}^n$ 中找到两个超平面:

$$f_+(x) = \omega_+ x + b_+ = 0, \quad f_-(x) = \omega_- x + b_- = 0, \quad (10)$$

在本文中, 分别称 $f_+(x)$ 和 $f_-(x)$ 为正类参数边缘超平面和负类参数边缘超平面。特别地, 如果给定训练集 T , $f_+(x)$ 和 $f_-(x)$ 分别决定正类和负类参数边缘超平面。通过合并两个参数边缘超平面, 平面, BP-TSVM 试图将数据分离成:

$$\omega_+^T x_i + b_+ \geq 0, \quad i = 1, 2, \dots, n_+, \quad (11)$$

$$\omega_-^T x_j + b_- \leq 0, \quad j = 1, 2, \dots, n_-, \quad (12)$$

相当于

$$1 \cdot f_+(x) \geq 0, \quad (-1) \cdot f_-(x) \leq 0. \quad (13)$$

通过引入有界弹球损失, 可以得到如下的一对新的优化问题:

$$\min_{\omega_+, b_+} \frac{1}{2} (\|\omega_+\|_2^2 + b_+^2) + \frac{\nu_1}{n_+} e_-^T (X_- \omega_+ + e_- b_+) + \frac{c_1}{n_+} \sum_{i=1}^{n_+} L_{BP}(x_i, y_i, f_+(x_i)), \quad (14)$$

和

$$\min_{\omega_-, b_-} \frac{1}{2} (\|\omega_-\|_2^2 + b_-^2) - \frac{\nu_2}{n_-} e_+^T (X_+ \omega_- + e_+ b_-) + \frac{c_2}{n_-} \sum_{j=1}^{n_-} L_{BP}(x_j, y_j, f_-(x_j)). \quad (15)$$

其中 $c_1, c_2 \geq 0$, $\nu_1, \nu_2 \geq 0$ 是惩罚参数。方便起见, 将损失函数中的 η 固定为 1 并吸收到 c_1, c_2 中, 即 $L_{BP^1}(u) = L_{BP}(u; \eta = 1)$, 令 $\omega_{\pm} = (\omega_{\pm}^T, b_{\pm})^T$, $X_{\pm} = (X_{\pm}, 1)$,

则优化目标可简化为:

$$\min_{\omega_+} \frac{1}{2} \|\omega_+\|_2^2 + \frac{\nu_1}{n_+} e_-^T X_- \omega_+ + \frac{c_1}{n_+} \sum_{i=1}^{n_+} L_{BP^1}(-(\omega_+^T x_i)), \quad (16)$$

和

$$\min_{\omega_-} \frac{1}{2} \|\omega_-\|_2^2 - \frac{\nu_2}{n_-} e_+^T X_+ \omega_- + \frac{c_2}{n_-} \sum_{j=1}^{n_-} L_{BP^1}(\omega_-^T x_j). \quad (17)$$

其中 $z_i = 0 - y_i f_+(x_i) = -\omega_+^T x_i$ 。可将目标函数改写为:

$$\min_{\omega_-} \frac{1}{2} \|\omega_-\|_2^2 - \frac{\nu_2}{n_-} e_+^T X_+ \omega_- + \frac{c_2}{n_-} \sum_{j=1}^{n_-} L_{BP^1}(\omega_-^T x_j). \quad (18)$$

$$\text{s.t.} \quad X_+ \omega_+ \geq 0 - \frac{\xi_+}{\tau_1}, \quad X_+ \omega_+ \leq 0 + \frac{\xi_+}{\lambda \tau_1}$$

其中 $\delta_+^k = (\delta_1^k, \delta_2^k, \dots, \delta_{nk}^k)^T$, ξ_+ 是松弛向量, 是非负调节参数, $c_1 > 0, \nu_1 > 0$ 是惩罚参数。拉格朗日函数为:

$$\begin{aligned} & \min_{\omega_-} \frac{1}{2} \|\omega_-\|_2^2 - \frac{\nu_2}{n_-} e_+^T X_+ \omega_- + \frac{c_2}{n_-} \sum_{j=1}^{n_-} L_{BP^1}(\omega_-^T x_j) \\ & + \alpha^T \left(-\frac{\xi_+}{\lambda} - X_+ \omega_+ \right) + \beta^T \left(-\frac{\xi_+}{\lambda \tau_1} + X_+ \omega_+ \right) \end{aligned} \quad (19)$$

利用拉格朗日函数对 ω_+ 和 ξ_+ 求偏导并令其为 0, 有:

$$\begin{aligned}\frac{\partial L(\omega_+, \xi_+, \alpha, \beta)}{\partial \omega_+} &= \omega_+ + \frac{v_1}{n_-} X_-^T e_- + \frac{c_1}{n_+} X_+^T \delta_k^+ - X_+^T \alpha + X_+^T \beta = 0 \\ \frac{\partial L(\omega_+, \xi_+, \alpha, \beta)}{\partial \xi_+} &= \frac{c_1}{n_+} e_+ - \frac{\alpha}{\lambda} - \frac{\beta}{\lambda \tau_1} = 0 \\ \alpha^T \left(-\frac{\xi_+}{\lambda} - X_+ \omega_+ \right) &= 0 \\ \beta^T \left(-\frac{\xi_+}{\lambda \tau_1} - X_+^+ \omega_+ \right) &= 0\end{aligned}\quad (20)$$

由(20)可知

$$\omega_+ = X_+^T \left(\alpha - \beta - \frac{c_1}{n_+} \delta_k^+ \right) - \frac{v_1}{n_-} X_-^T e_-, \frac{c_1}{n_+} e_+ = \frac{\alpha}{\lambda} + \frac{\beta}{\lambda \tau_1}. \quad (21)$$

将(21)代入拉格朗日函数, 并记 $u = \alpha - \beta - \delta_k^+$, 可得到目标函数的对偶问题:

$$\begin{aligned}\min_u \quad & \frac{1}{2} u^T X_+ X_+^T u - \frac{v_1}{n_-} X_- X_+^T u \\ \text{s.t.} \quad & -\frac{c_1}{n_+} \delta_k^+ - \frac{c_1}{n_+} \lambda \tau_1 e_+ \leq u \leq -\frac{c_1}{n_+} \delta_k^+ - \frac{c_1}{n_+} \lambda e_+.\end{aligned}\quad (22)$$

其中 $q_j = 0 - y_j f_-(x_j) = \omega_-^T x_j$ 。将目标函数改写为:

$$\begin{aligned}\min_{\omega_-} \quad & \frac{1}{2} \|\omega_-\|^2 - \frac{v_2}{n_+} e_+^T X_+ \omega_- + \frac{c_2}{n_-} e_-^T \xi_- + \frac{c_2}{n_-} \delta_k^- X_-^T \omega_- \\ \text{s.t.} \quad & -X_- \omega_- \geq 0 - \frac{\xi_-}{\lambda}, \quad -X_- \omega_- \leq 0 + \frac{\xi_-}{\lambda \tau_2}.\end{aligned}\quad (23)$$

其中 $\delta_k^- = (\delta_1 k^-, \delta_2 k^-, \dots, \delta_{n-k}^-)^T$, ξ_- 是松弛向量, τ_2 是非负调节参数, $c_2 > 0, v_2 > 0$ 是惩罚参数。同时对偶问题如下:

$$\begin{aligned}\min_v \quad & \frac{1}{2} v^T X_- X_-^T v + \frac{v_2}{n_+} X_+ X_-^T v \\ \text{s.t.} \quad & -\frac{c_2}{n_-} \delta_k^- - \frac{c_2}{n_-} \lambda e_- \leq v \leq -\frac{c_2}{n_-} \delta_k^- + \frac{c_2}{n_-} \lambda \tau_2 e_+.\end{aligned}\quad (24)$$

变量 u 可通过下式计算

$$\omega_- = X_-^T v + \frac{v_2}{n_+} X_+^T e_+. \quad (25)$$

通过求解二次规划问题, 可以得到 u , u , 从而得到 ω_+ , ω_- 。最后, 我们可以得到 BP-TSVM 的决策函数为:

$$f(x) = \text{sgn} \left(\frac{\omega_+ \cdot x + b_+}{\|\omega_+\|} + \frac{\omega_- \cdot x + b_-}{\|\omega_-\|} \right). \quad (26)$$

3.3. 非线性

利用核技巧将线性 BP-TSVM 扩展到非线性情况。将训练集通过非线性映射 $\phi(x): x \rightarrow \phi(x) \in H$ ，将映射到希尔伯特空间中，在此选择合适的核函数 $K(x, x^T) = (\phi(x), \phi(x)^T)$ 。结合 CCCP 和核技巧，非线性情况下，则非线性 BP-TSVM 的优化目标函数为：

$$\min_{\omega_+} \frac{1}{2} \|\omega_+\|^2 + \frac{v_1}{n_-} e_-^T \phi(x_-) \omega_+ + \frac{c_1}{n_+} e_+^T \xi_+ + \frac{c_1}{n_+} \delta_k^{+T} \phi(x_+) \omega_+, \quad (27)$$

$$\text{s.t. } \phi(x_+) \omega_+ \geq 0 - \frac{\xi_+}{\tau_1}, \quad \phi(x_+) \omega_+ \leq 0 + \frac{\xi_+}{\lambda \tau_1}. \quad (28)$$

和

$$\min_{\omega_-} \frac{1}{2} \|\omega_-\|^2 - \frac{v_2}{n_+} e_+^T \phi(x_+) \omega_- + \frac{c_2}{n_-} e_-^T \xi_- + \frac{c_2}{n_-} \delta_k^{-T} \phi(x_-) \omega_-$$

$$\text{s.t. } \phi(x_-) \omega_- \geq 0 - \frac{\xi_-}{\lambda}, \quad \phi(x_-) \omega_- \leq 0 + \frac{\xi_-}{\lambda \tau_2}. \quad (29)$$

其中 $c_1, c_2 > 0, v_1, v_2 > 0$ 是惩罚参数， ξ_+, ξ_- 是松弛向量， τ_1, τ_2 是非负调节参数。类似地，我们可以推导出(27)~(28)的对偶问题：

$$\min_u \frac{1}{2} u^T K(X_+, X_+) u - \frac{v_1}{n_-} K(X_-, X_+) u$$

$$\text{s.t. } -\frac{c_1}{n_+} \delta_k^+ - \frac{c_1}{n_+} \lambda \tau_1 e_+ \leq u \leq -\frac{c_1}{n_+} \delta_k^+ - \frac{c_1}{n_+} \lambda e_+. \quad (30)$$

和

$$\min_v \frac{1}{2} v^T K(X_-, X_-) v + \frac{v_2}{n_+} K(X_+, X_-) v$$

$$\text{s.t. } -\frac{c_2}{n_-} \delta_k^- - \frac{c_2}{n_-} \lambda e_- \leq v \leq -\frac{c_2}{n_-} \delta_k^- + \frac{c_2}{n_-} \lambda \tau_2 e_+. \quad (31)$$

与线性情况类似，在优化两个二次规划问题(30)~(31)之后，非线性 BP-TSVM 的决策函数可构造为(26)。

4. 在线购物者购买意向预测分析研究

4.1. 数据来源

本研究所使用的在线购物者购买意向(Online Shoppers Purchasing Intention)数据集来源于 UCI Machine Learning Repository，该数据集包含 12,330 条会话记录，每条记录代表来自不同用户的一次独立访问，数据采集周期为一年，以减少季节性促销或特定节日的偏倚影响。数据集中共有 17 个特征变量与 1 个目标变量，涵盖用户在电商平台上的访问行为、站内交互特征及部分外部环境信息。目标变量为二分类标签 Revenue，取值 True 表示该会话产生购买，False 表示未产生购买，约 1.5 成的样本为购买类别，其余为未购买类别，呈现显著的类别不平衡特征，数据集信息见表 1。

Table 1. Variables in the online shoppers' purchase intention dataset
表 1. 在线购物者购买意向数据集变量说明

变量名称	变量说明
数值型特征(10 个)	
Administrative	访问管理类页面的次数(如账户信息、登录等)
Administrative_Duration	管理类页面的总停留时间(秒)
Informational	访问信息类页面的次数(如网站政策、联系方式等)信息类页面的总停留时间(秒)
Informational_Duration	访问产品相关页面的次数
ProductRelated	
ProductRelated_Duration	产品页面的总停留时间(秒)
BounceRates	跳出率(仅访问该页面的会话比例)
ExitRates	退出率(以该页面结束的会话比例)
PageValues	页面平均价值(页面对购买的贡献度)
SpecialDay	特殊日期接近度(0~1, 越接近节日值越高)
类别型特征(7 个)	
Month	访问发生的月份
OperatingSystems	用户使用的操作系统类型
Browser	用户使用的浏览器类型
Region	用户所在地区
TrafficType	访问来源类型(如直接访问、广告、推荐等)
VisitorType	访客类型(新访客或回访访客)
Weekend	是否为周末访问
目标变量	
Revenue	会话是否产生购买(True/False)

4.2. 在线购物者购买意向预测模型建立

本节中进行实证研究, 将提出的 BP-TSVM 应用于在线购物者购买意向预测模型, 实验都在 R 4.3.1 中进行, 基于系统配置 Intel(R) Core(TM) i7-10700 CPU (2.90GHz)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 TSVM、TPMSVM, 设置 $C = \{2^{-8}, 2^{-7}, \dots, 2^7, 2^8\}$, 设置 Pin-TSVM 和 BP-TSVM 中 τ 的取值范围为 $\tau = \{0.1, 0.2, 0.5, 1\}$ 。

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 10 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 2 所示。提出的模型提高了预测准确率, 将其应用于在线购物者购买意向预测模型可以带来更好的效果。

Table 2. Prediction outcomes of the Dataset
表 2. 数据集结果预测

	BP-TSVM	Pin-TSVM	TPMSVM	TSVM
Acc.	0.9149	0.8741	0.7922	0.8823
Sd.	0.0136	0.0144	0.0545	0.0523

5. 结论

本文针对电商平台的在线购买者行为数据, 构建并验证了购买意向预测模型。在充分分析 UCI 在线购物者购买意向数据集的基础上, 本文提出了基于有界分位损失的孪生支持向量机(BP-TSVM)模型, 并将其与经典孪生支持向量机(TSVM)、Pin-TSVM 以及最小二乘 TPMSVM 进行了对比实验。实验结果表明, BP-TSVM 在预测精度和鲁棒性方面均优于其他对比模型, 尤其在处理含噪声的电商用户行为数据时表现更加稳定。此外, 在针对如电商平台的在线购买者行为数据这种类别不平衡的数据, 预测效果也较好。

本研究的结论表明, 在电商背景下, 利用用户访问行为、页面停留时长、跳出率、页面价值及节日接近度等多维特征, 可以有效预测用户的购买意向。BP-TSVM 模型不仅能够提升购买意向预测的准确性, 还为电商平台在精准营销、个性化推荐、用户分层和转化率优化等方面提供了有力的技术支持。此外, 本研究的方法和实验结果为未来在高维、噪声复杂的电商行为数据中进行分类建模提供了可行的参考与实践基础。

未来工作可从以下几个方向展开: (1) 结合特征选择方法提升模型可解释性; (2) 探索 BP-TSVM 在其他不平衡或多分类金融数据集中的扩展性; (3) 将该模型嵌入电商购买者购买意愿预测中, 进行更大规模实证研究。

参考文献

- [1] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. <https://doi.org/10.1023/a:1022627411411>
- [2] Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Nédellec, C. and Rouveirol, C., Eds., *Machine Learning: ECML-98*, Springer Berlin Heidelberg, 137-142. <https://doi.org/10.1007/bfb0026683>
- [3] Moghaddam, B. and Yang, M.H. (1998) Face Recognition Using Support Vector Machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1066-1073.
- [4] Liu, C. and Wechsler, H. (2003) Face Recognition Using Support Vector Machines. *IEEE Transactions on Neural Networks*, **14**, 195-200.
- [5] Burges, C.J.C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, **2**, 121-167. <https://doi.org/10.1023/a:1009715923555>
- [6] Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992) A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, 27-29 July 1992, 144-152. <https://doi.org/10.1145/130385.130401>
- [7] Osuna, E., Freund, R. and Girosi, F. (1997) Training Support Vector Machines: An Application to Face Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, 17-19 June 1997, 130-136.
- [8] Platt, J.C. (1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsoft Research.
- [9] Shen, X., Zhang, Y. and Chen, X. (2017) Support Vector Machine with Noisy Features: Robustness Analysis and Feature Selection. *IEEE Transactions on Neural Networks and Learning Systems*, **28**, 2335-2346.
- [10] Huang, J., Zhang, L. and Metaxas, D. (2013) Support Vector Machines with Pinball Loss Function. *Pattern Recognition*, **46**, 1552-1563.
- [11] Wu, T. and Liu, C. (2007) Robust Truncated Hinge Loss Support Vector Machines. *Journal of Machine Learning Research*, **8**, 2301-2327.
- [12] Tang, Z., Chen, Y. and Li, J. (2021) Valley Loss for Robust Multi-Class Classification. *Neurocomputing*, **453**, 260-272.
- [13] 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>
- [14] 谢平, 邹传伟, 刘海二. 互联网金融模式研究[J]. 金融研究, 2012(12): 11-22.

- [15] 孟陆, 刘凤军, 陈斯允, 段坤. 我可以唤起你吗——不同类型直播网红信息源特性对消费者购买意愿的影响研究[J]. 管理世界, 2020, 36(4): 142-156.
- [16] 江积海, 李琴. 平台型企业商业模式创新中“属性-功能”认知演化机理: 基于海尔和苹果的案例分析[J]. 管理评论, 2016, 28(7): 249-257.