Statistics and Application 统计学与应用, 2025, 14(8), 55-62
Published Online August 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.148215

Hans汉斯

基于岭惩罚与修正的SCAD惩罚的软间隔SVM分类模型与算法及其应用

扶庆阳，王钧传，张　鸿，高林庆*

河北工程大学数理科学与工程学院，河北　邯郸

收稿日期：2025年7月6日；录用日期：2025年7月26日；发布日期：2025年8月7日

摘　要

基于岭惩罚与传统的SCAD惩罚的软间隔SVM分类算法已成功应用于医学诊断等领域。然而传统的SCAD惩罚只考虑样本信息的影响，未考虑到先验信息的影响，故此，基于岭惩罚与传统的SCAD惩罚的软间隔SVM分类算法有一定的局限性。修正的SCAD惩罚同时考虑了样本信息与先验信息的影响，是传统的SCAD惩罚的重要拓广，目前未发现基于岭惩罚与修正的SCAD惩罚的软间隔SVM分类算法的研究。基于此，本文首先将岭惩罚与修正的SCAD惩罚相结合，并与软间隔SVM分类算法融合，构建基于岭惩罚与修正的SCAD惩罚的软间隔SVM分类模型。然后，引入AIC和BIC信息准则求解参数。最后，通过心脏病诊断实例验证了所提模型与算法具有更高的灵敏度、特异度和分类能力。

关键词

岭惩罚，修正的SCAD惩罚，软间隔支持向量机，分类算法

Soft Margin SVM Classification Model and Algorithm Based on Ridge Penalty and Modified SCAD Penalty and Its Applications

Qingyang Fu, Junchuan Wang, Hong Zhang, Linqing Gao*

School of Mathematics and Physics, Hebei University of Engineering, Handan Hebei

Received: Jul. 6th, 2025; accepted: Jul. 26th, 2025; published: Aug. 7th, 2025

Abstract

Soft margin SVM classification algorithms based on ridge penalty and traditional SCAD penalty have

*通讯作者。

文章引用: 扶庆阳, 王钧传, 张鸿, 高林庆. 基于岭惩罚与修正的 SCAD 惩罚的软间隔 SVM 分类模型与算法及其应用[J]. 统计学与应用, 2025, 14(8): 55-62. DOI: 10.12677/sa.2025.148215

been successfully applied in fields such as medical diagnosis. However, traditional SCAD penalty only considers the influence of the sample information and does not take into account the influence of prior information. Therefore, soft margin SVM classification algorithms based on ridge penalty and traditional SCAD penalty have certain limitations. The modified SCAD penalty simultaneously considers the influence of both the sample information and prior information, representing an important extension of the traditional SCAD penalty. To date, no research has been found on soft margin SVM classification algorithms based on ridge penalty and modified SCAD penalty. Based on this, this paper first combines ridge penalty with modified SCAD penalty and integrates them with the soft margin SVM classification algorithm to construct a soft margin SVM classification model based on ridge penalty and modified SCAD penalty. Then, AIC and BIC information criteria are introduced to solve the parameters. Finally, the proposed algorithm is validated through a cardiac diagnosis example, demonstrating higher sensitivity, specificity, and classification capability.

Keywords

Ridge Penalty, Modified SCAD Penalty, Soft Margin SVM, Classification Algorithm

1. 引言

岭惩罚由 Tikhonov 于 20 世纪 40 年代提出[1]，能够有效缓解特征间的多重共线性问题。软间隔支持向量机(Support Vector Machine, SVM)由 Cortes 和 Vapnik 提出[2] [3]，通过引入松弛变量及相应的惩罚项，允许部分样本位于分类边界之外，即容忍一定程度的分类错误，使其能够适用于非线性可分的数据集。岭惩罚与软间隔 SVM 融合在面对多重共线性特征时，对特征系数进行均匀收缩，降低模型的方差，提升模型泛化能力，提高分类精度。但是，岭惩罚与软间隔 SVM 融合不能进行自动特征选择且同时完成高维特征数据分类任务。平滑剪切绝对偏差(Smoothly Clipped Absolute Deviation, SCAD)惩罚[4]旨在克服岭惩罚不会产生稀疏解问题，在较大系数时减小惩罚力度，从而减小估计偏差，同时实现特征稀疏选择。Natalia Becker [5]提出了一种融合岭惩罚与 SCAD 惩罚的软间隔 SVM 分类算法，该分类算法为医疗诊断任务提供了灵活且稳健的工具，尤其适用于如微阵列数据集等高维生物医学数据的分类与特征筛选。然而，传统的 SCAD 惩罚与软间隔 SVM 融合未考虑到先验信息对调谐参数的影响，这可能导致无法选择出最优的调谐参数，并且对于较大的调谐参数 λ，传统的 SCAD 惩罚趋于平缓，降低了其惩罚效果。修正的 SCAD 惩罚是 Ng 和 Yu [6]提出的一种结合了先验信息与样本信息的惩罚，并引入新定义的 AIC 的 BIC 克服了传统 SCAD 惩罚的缺陷，可以在保持良好特征选择能力和分类性能的同时，提高调谐参数选择的准确性和模型的泛化能力。

为了有效克服上述基于岭惩罚与传统的 SCAD 惩罚的软间隔 SVM 分类算法的局限性，本文将岭惩罚和修正的 SCAD 惩罚组合，并与软间隔 SVM 融合，构建了岭惩罚与修正的 SCAD 的软间隔 SVM 分类算法，并应用于心脏病诊断实例中。

2. 预备知识

2.1. 岭惩罚

软间隔 SVM 的惩罚项采用 L_2 范数即岭惩罚[1]：

$$pen_{\lambda_1}(\boldsymbol{\omega}) = \lambda_1 \|\boldsymbol{\omega}\|_2^2 = \lambda_1 \sum_{i=1}^{p} \omega_i^2 . \tag{1}$$

岭惩罚通过对系数进行收缩以控制其方差。

2.2. 修正的 SCAD 惩罚

传统 SCAD 惩罚是 Fan 和 Li [4]首先提出的一种非凸惩罚函数。Zhang 等[7]将软间隔 SVM 算法与传统 SCAD 惩罚相结合进行特征选择。对于单个系数 ω_i 的 SCAD 惩罚函数定义为：

$$p_{\mathrm{SCAD}(\lambda_2)}(\omega_i) = \begin{cases} \lambda_2 |\omega_i| & |\omega_i| \le \lambda_2 \\ -\dfrac{|\omega_i|^2 - 2a\lambda_2|\omega_i| + \lambda^2}{2(a-1)} & \lambda_2 < |\omega_i| \le a\lambda_2 \\ \dfrac{(a+1)\lambda_2^2}{2} & |\omega_i| > a\lambda_2 \end{cases} \tag{2}$$

其中 $\omega_i, i = 1,2,\cdots,p$ 是定义超平面的系数，$a > 0$ 和 $\lambda_2 > 0$ 是调谐参数。针对软间隔 SVM 算法的传统 SCAD 惩罚项具有以下形式：

$$pen_{\lambda_2}(\boldsymbol{\omega}) = \sum_{i=1}^{p} P_{SCAD(\lambda_2)}(\omega_i). \tag{3}$$

传统 SCAD 惩罚对于 λ_2 和 $a\lambda_2$ 之间的系数 ω_i 采用二次惩罚(非线性缓惩)，对于小系数 ω_i，传统 SCAD 惩罚采用与 L_1 范数相同的惩罚，而对于较大的系数 ω_i，传统 SCAD 采用常数惩罚。

相对传统 SCAD 惩罚只结合了样本进行特征选择，修正的 SCAD 惩罚是一种结合了先验信息与样本信息的惩罚函数，它通过引入新定义的 AIC 和 BIC 克服了传统 SCAD 惩罚函数选取调谐参数的缺陷，对于单个系数 ω_i 的修正的 SCAD 惩罚定义为：

$$P_{\lambda_2}(\omega_i) = \begin{cases} \lambda_2 k^{-1}\left[(1+|\omega_i|)^k - 1\right], & |\omega_i| \le \lambda_2, \\ \dfrac{-(1+\lambda_2)^{k-1}}{2(a-1)}\left[\omega_i^2 - 2a\lambda_2|\omega_i| + (2a-1)\lambda_2^2\right] + \dfrac{\lambda_2}{k}\left[(1+\lambda_2)^k - 1\right], & \lambda_2 < |\omega_i| \le a\lambda_2, \\ \dfrac{a-1}{2}(1+\lambda_2)^{k-1}\lambda_2^2 + \dfrac{\lambda_2}{k}\left[(1+\lambda_2)^k - 1\right], & |\omega_i| > a\lambda_2, \end{cases} \tag{4}$$

其中 $a > 2$ 和 $\lambda_2 > 0$，k 取决于样本量，修正的 SCAD 惩罚函数调谐参数可以通过新定义的 AIC 和 BIC 信息标准进行选取。针对软间隔 SVM 算法的修正的 SCAD 惩罚项具有以下形式：

$$pen_{\lambda_2}(\boldsymbol{\omega}) = \sum_{i=1}^{p} P_{\mathrm{mod}\,SCAD(\lambda_2)}(\omega_i) \tag{5}$$

2.3. 软间隔支持向量机

SVM 的硬间隔分类器存在一定的局限性[2] [3]，为此，软间隔 SVM 引入松弛变量，允许在一定程度上违反间隔约束，允许某些数据点位于间隔的错误一侧。松弛变量 $\xi_i \ge 0,\ i = 1,2,\cdots,n$ 被定义其错误分类点与对应的边距距离，对于边缘正确一侧的数据点 $\xi_i = 0$，对于边缘内部的数据点 $0 < \xi_i \le 1$，并对错误的分类数据点 $\xi_i > 1$。用代价参数 C 对非零 ξ_i 之和进行惩罚，然后将其添加到最小化问题的优化函数惩罚中：

$$\min_{b,\omega}\frac{1}{2}\|\boldsymbol{\omega}\|_2^2+C\sum_{i=1}^{n}\xi_i,$$
$$\text{s.t. } \xi_i\geq 0,\ \ i=1,2,\cdots,n, \tag{6}$$
$$y_i\left(b+\boldsymbol{\omega}\cdot x_i\right)\geq 1-\xi_i\quad i=1,2,\cdots,n.$$

上述优化问题称为软间隔 SVM，其中代价参数 C 是一个依赖数据样本大小的调节参数，它控制着最小化超平面的系数和训练数据集的正确分类之间的平衡。代价参数 C 常通过交叉验证进行选择。并且可以通过使用凸优化技术，即拉格朗日乘子法来求解[8]。凸优化技术为超平面参数 $\boldsymbol{\omega}$ 和 b 提供了解决方法：

$$\hat{\boldsymbol{\omega}}=\sum_{i=1}^{n}\alpha_i y_i x_i, \tag{7}$$

其中 $\alpha_i\geq 0,\ i=1,2,\cdots,n$ 是 Lagrange 乘子，α_i 为正的数据点称为支持向量。所有位于其边距正确一侧的数据点 $\alpha_i=0$，因此，它们对超平面没有任何影响，可以重写上述方程如下：

$$\hat{\boldsymbol{\omega}}=\sum_{s\in S}\alpha_s y_s x_s \tag{8}$$

其中支持向量 S 的指标集由 $S=\{i:\alpha_i>0\}$，对于任意 i，当 $\alpha_i>0$ 时，由 $y_i\left(\hat{\boldsymbol{\omega}}\cdot x_i+\hat{b}\right)=1-\xi_i$ 可计算出系数 $\hat{b}$。在实际应用中，对于 $\hat{b}$ 所有解的均值被用于数值稳定性。

3. 基于岭惩罚与修正的 SCAD 惩罚的软间隔 SVM 分类算法

3.1. 基于岭惩罚与修正的 SCAD 惩罚的软间隔 SVM 分类模型

如第一节引言所述，基于岭惩罚与传统 SCAD 的软间隔 SVM 分类算法具有一定的局限性。为了有效地克服此局限性，本节首先将岭惩罚与修正的 SCAD 惩罚组合成新的惩罚，新的组合惩罚函数如下：

$$pen_\lambda\left(\boldsymbol{\omega}\right)=\lambda_1\|\boldsymbol{\omega}\|_2^2+\sum_{i=1}^{p}P_{\mathrm{mod}\,SCAD(\lambda_2)}\left(\omega_i\right), \tag{9}$$

其中 $\lambda_1,\lambda_2\geq 0$，为调谐参数。可以看出，当岭惩罚的调谐参数收敛到零($\lambda_1\to 0$)，组合惩罚提供了稀疏性、连续性和渐近正态性。将新的组合惩罚与软间隔支持向量机融合，构建基于岭惩罚与修正的 SCAD 惩罚的软间隔 SVM 分类模型。首先给定目标函数如下：

$$\min_{\omega,b}\frac{1}{n}\sum_{i=1}^{n}\left[1-y_i f\left(x_i\right)\right]_+ +\lambda_1\|\boldsymbol{\omega}\|_2^2+\sum_{j=1}^{p}P_{\mathrm{mod}\,SCAD(\lambda_2)}\left(\omega_j\right), \tag{10}$$

其中 $\lambda_1,\lambda_2\geq 0$，为调谐参数。同时为应对高惩罚或约束下 AIC 和 BIC 的局限性，引入了一个受约束的信息准则，AIC 和 BIC 表达式如下：

$$\mathrm{AIC}=-2\log L\left(\boldsymbol{\omega}\middle|Y\right)+2\mathrm{eff}\left(\lambda_2\right), \tag{11}$$

$$\mathrm{BIC}=-2\log L\left(\boldsymbol{\omega}\middle|Y\right)+\mathrm{eff}\left(\lambda_2\right)\cdot\log n, \tag{12}$$

其中 $L\left(\boldsymbol{\omega}\middle|Y\right)$ 是似然函数，n 是样本量，其中 $\mathrm{eff}\left(\lambda_2\right)$ 定义如下：

$$\mathrm{eff}\left(\lambda_2\right)=\mathrm{trace}\left(\boldsymbol{H}^R\left(\lambda_2\right)\right), \tag{13}$$

其中 $\boldsymbol{H}^R\left(\lambda_2\right)$ 是考虑惩罚和约束的帽子矩阵，$\boldsymbol{H}^R\left(\lambda_2\right)$ 的定义如下：

$$\boldsymbol{H}^R\left(\lambda_2\right)=\boldsymbol{X}_n\left(\boldsymbol{A}_n^{-1}-\boldsymbol{A}_n^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{A}_m^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{A}_n^{-1}\right)\boldsymbol{X}_n^{\mathrm{T}}, \tag{14}$$

其中 $\boldsymbol{X}_n$ 是设计矩阵，$\boldsymbol{A}_n = \boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{X}_n + n\lambda_2\boldsymbol{I}_p$ 是惩罚设计矩阵，$\boldsymbol{R}$ 是约束矩阵，$\boldsymbol{I}_p$ 是 $p\times p$ 的单位矩阵。最后，在给定约束条件下，合理选择 AIC 和 BIC 实现了模型拟合与稀疏性的平衡。该准则有效地惩罚小系数，同时最小化对大系数的影响。

3.2. 基于岭惩罚与修正的 SCAD 的软间隔 SVM 分类算法

上述 3.1 构建的基于岭惩罚与修正的 SCAD 惩罚的软间隔 SVM 分类模型的求解算法，类似于岭惩罚与传统的 SCAD 惩罚的软间隔 SVM 分类算法[5]。为了方便起见，将修正的 SCAD 惩罚从 $P_{\mathrm{mod}\,SCAD(\lambda_2)}(\omega_i)$ 重新命名为 $P_{\lambda_2}(\omega_i)$。相应地，惩罚的一阶导数记为 $P'_{\lambda_2}(\cdot)$，即

$$P'_{\lambda_2}(\omega_i)=\begin{cases}\lambda_2\left(1+|\omega_i|\right)^{k-1}\mathrm{sgn}(\omega_i), & |\omega_i|\le\lambda_2\\ \dfrac{\left(1+\lambda_2\right)^{k-1}}{a-1}\left(a\lambda_2-|\omega_i|\right)\mathrm{sgn}(\omega_i) & \lambda_2<|\omega_i|\le a\lambda_2.\\ 0 & |\omega_i|>a\lambda_2\end{cases} \tag{15}$$

则目标函数式(10)表示为：

$$A(b,\boldsymbol{\omega}):=\frac{1}{n}\sum_{i=1}^{n}\left[1-y_i f(x_i)\right]_+ + \lambda_1\|\boldsymbol{\omega}\|_2^2 + \sum_{j=1}^{p}P_{\lambda_2}(\omega_i). \tag{16}$$

对于每一个 i，损失项可以根据

$$\left[1-y_i\left(b+\boldsymbol{\omega}\cdot x_i\right)\right]_+ = \frac{1-y_i\left(b+\boldsymbol{\omega}\cdot x_i\right)}{2}+\frac{\left|1-y_i\left(b+\boldsymbol{\omega}\cdot x_i\right)\right|}{2},$$

给定接近 $A(b,\omega)$ 的最小值的初值 (b_0,ω_0)，考虑如下局部近似表达式

$$\left|y_i-\left(b+\boldsymbol{\omega}\cdot x_i\right)\right|\approx\frac{1}{2}\frac{\left\{y_i-\left(b+\boldsymbol{\omega}\cdot x_i\right)\right\}^2}{\left\{y_i-\left(b_0+\boldsymbol{\omega}_0\cdot x_i\right)\right\}}+\frac{1}{2}\left|y_i-\left(b_0+\boldsymbol{\omega}_0\cdot x_i\right)\right|,$$

当 ω_{j0} 接近于零时，设 $\hat{\omega}_j=0$，否则使用修正的 SCAD 惩罚项的近似表达式：

$$P_{\lambda_2}(\omega_j)\approx P_{\lambda_2}(\omega_{j0})+\frac{1}{2}\frac{P'_{\lambda_2}(\omega_{j0})}{|\omega_{j0}|}\left(\omega_j^2-\omega_{j0}^2\right),$$

其中，由于修正的 SCAD 惩罚项的对称性，使用 $|\omega_j|$ 代替 ω_j。

$A(b,\boldsymbol{\omega})$ 的局部二次近似具有以下形式：

$$\begin{aligned}A(b,\boldsymbol{\omega})\approx{}&\frac{1}{2}-\frac{1}{2n}\sum_{i=1}^{n}y_i\left(b+\boldsymbol{\omega}\cdot x_i\right)+\frac{1}{4n}\sum_{i=1}^{n}\left|y_i-\left(b_0+\boldsymbol{\omega}_0\cdot x_i\right)\right|\\&+\frac{1}{4n}\sum_{i=1}^{n}\frac{\left[y_i-\left(b+\boldsymbol{\omega}\cdot x_i\right)\right]^2}{\left|y_i-\left(b_0+\boldsymbol{\omega}_0\cdot x_i\right)\right|}+\sum_{j=1}^{p}\lambda_1\omega_j^2+\sum_{j=1}^{p}P_{\lambda_2}(\omega_{j0})+\sum_{j=1}^{p}\frac{P'_{\lambda_2}(\omega_{j0})}{2|\omega_{j0}|}\left(\omega_j^2-\omega_{j0}^2\right).\end{aligned}$$

通过对 $A(b,\boldsymbol{\omega})$ 关于 $\boldsymbol{\omega}$ 和 b 的最小化，可以省略不含优化参数 $\boldsymbol{\omega}$ 和 b 的项(因为常数的导数为零)

$$\begin{aligned}A(b,\boldsymbol{\omega})\approx{}&-\frac{1}{2n}\sum_{i=1}^{n}y_i\left(b+\boldsymbol{\omega}\cdot x_i\right)+\frac{1}{2n}\sum_{i=1}^{n}\frac{y_i\left(b+\boldsymbol{\omega}\cdot x_i\right)}{\left|y_i-\left(b_0+\boldsymbol{\omega}_0\cdot x_i\right)\right|}\\&+\frac{1}{4n}\sum_{i=1}^{n}\frac{\left(b+\boldsymbol{\omega}\cdot x_i\right)^2}{\left|y_i-\left(b_0+\boldsymbol{\omega}_0\cdot x_i\right)\right|}+\sum_{j=1}^{p}\lambda_1\omega_j^2+\sum_{j=1}^{p}\frac{P'_{\lambda_2}(\omega_{j0})}{2|\omega_{j0}|}\cdot\omega_j^2.\end{aligned}$$

为了将方程写成矩阵形式，定义如下：

$$\boldsymbol{y}=\left[y_1,y_2,\cdots,y_n\right]^{\mathrm{T}},$$

$$\boldsymbol{\omega}=\left[\omega_1,\omega_2,\cdots,\omega_p\right]^{\mathrm{T}},$$

$$\boldsymbol{\varepsilon}=\left[\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\right]^{\mathrm{T}},\text{其中}\quad \varepsilon_i=y_i-\left(b_0+\omega_0\cdot x_i\right).$$

此外，定义矩阵 $\boldsymbol{X}=\left[1,x_1,x_2,\cdots,x_p\right]$，其中 $\mathbf{1}$ 是长度为 n 的全 1 向量，$\boldsymbol{x}_j$ 是第 j 个输入向量。设

$$\boldsymbol{r}=\left[\frac{y_1}{|\varepsilon_1|},\frac{y_2}{|\varepsilon_2|},\cdots,\frac{y_n}{|\varepsilon_n|}\right]^{\mathrm{T}},\quad \boldsymbol{D}_0=\frac{1}{2n}\mathrm{diag}\left[\frac{1}{|\varepsilon_1|},\frac{1}{|\varepsilon_2|},\cdots,\frac{1}{|\varepsilon_n|}\right],$$

$$\boldsymbol{Q}_1=\mathrm{diag}\left[0,\frac{p'_{\lambda_2}\left(\omega_{10}\right)}{|\omega_{10}|},\frac{p'_{\lambda_2}\left(\omega_{20}\right)}{|\omega_{20}|},\cdots,\frac{p'_{\lambda_2}\left(\omega_{d0}\right)}{|\omega_{d0}|}\right],\quad \boldsymbol{Q}_2=\mathrm{diag}\left[0,2\lambda_2,\cdots,2\lambda_2\right],$$

$$\boldsymbol{P}=\frac{1}{2n}\left(\boldsymbol{y}+\boldsymbol{r}\right)^{\mathrm{T}}\boldsymbol{X}\ \text{且}\ \boldsymbol{Q}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}_0\boldsymbol{X}+\boldsymbol{Q}_1+\boldsymbol{Q}_2.$$

因此，最小化 $A\left(b,\omega\right)$ 等价于最小化二次函数

$$\tilde{A}\left(b,\boldsymbol{\omega}\right)=\frac{1}{2}\begin{pmatrix}b\\ \boldsymbol{\omega}\end{pmatrix}^{\mathrm{T}}\boldsymbol{Q}\begin{pmatrix}b\\ \boldsymbol{\omega}\end{pmatrix}-\boldsymbol{P}\begin{pmatrix}b\\ \boldsymbol{\omega}\end{pmatrix}. \tag{17}$$

上述解满足线性方程组

$$\boldsymbol{Q}\begin{pmatrix}\hat{b}\\ \hat{\boldsymbol{\omega}}\end{pmatrix}=\boldsymbol{P}. \tag{18}$$

基于岭惩罚与修正的 SCAD 的软间隔 SVM 分类算法通过以下迭代算法实现：

步骤 1：设 $k=1$，初始化 L_2 支持向量机指定初始值 $\left(b^{(1)},\boldsymbol{\omega}^{(1)}\right)$；

步骤 2：存储第 k 次迭代的解：$\left(b_0,\boldsymbol{\omega}_0\right)=\left(b^{(k)},\boldsymbol{\omega}^{(k)}\right)$；

步骤 3：通过求解 $\boldsymbol{Q}\begin{pmatrix}\hat{b}\\ \hat{\boldsymbol{\omega}}\end{pmatrix}=\boldsymbol{P}$ 最小化 $A\left(b,\boldsymbol{\omega}\right)$，并将解表示为 $\left(b^{(k+1)},\boldsymbol{\omega}^{(k+1)}\right)$；

步骤 4：若 $\left|A\left(b^{(k+1)},\boldsymbol{\omega}^{(k+1)}\right)-A\left(b^{(k)},\boldsymbol{\omega}^{(k)}\right)\right|<10^{-3}$，则 $\left(b^{(k+1)},\boldsymbol{\omega}^{(k+1)}\right)$ 为最优解；否则，令 $k=k+1$，转到步骤 2。

4. 数值实验

本节针对基于岭惩罚和 Lasso 惩罚[9]的软间隔 SVM (SVM-L_2-L_1)、基于岭惩罚和 SCAD 惩罚的软间隔 SVM (SVM-L_2-SCAD)、和基于岭惩罚和修正的 SCAD 惩罚的软间隔 SVM (SVM-L_2-modSCAD)等三种分类模型和算法进行数值实验。实验数据来源于 UCI 机器学习库的克利夫兰心脏病数据集[10]，数据集包含 14 个特征变量，包括年龄、性别、心绞痛类型、静息血压、血清胆固醇、最大心率、心电图结果等，目标变量为是否患有心脏病(表 1)。

实验过程。首先，数值实验根据目标变量是否患有心脏病将数据按照等比例分划分到训练集和测试集，训练集和测试集比例为 8:2。其次，设定一个大的调节参数区间，来选取最优调谐参数，以确保调谐参数不陷入局部最优，其中 SVM-L_2-SCAD 的两个调谐参数在选取的时候遵循的原则也不相同，选取调谐参数 λ_1 和 λ_2 分别采用网格搜索法和给定的 AIC 与 BIC 进行分别选择。对于 SVM-L_2-modSCAD 中的

Table 1. Partial data table of the Cleveland heart disease dataset
表 1. 克利夫兰心脏病数据集部分数据表

序号	年龄	性别	心绞痛类型	静息血压	血清胆固醇	最大心率	心电图结果	是否患有心脏病	...
1	63	1	3	145	233	150	0	1	...
2	37	1	2	130	250	187	0	1	...
3	41	0	1	130	204	172	2	1	...
4	62	0	0	140	268	160	0	0	...
5	63	1	0	130	254	147	1	0	...
6	53	1	0	140	203	155	0	0	...

注：性别：0 为女性，1 为男性；心绞痛类型：1 为典型的心绞痛，2 为非典型心绞痛，3 为非心绞痛，4 为无症状；心电图结果为最高运动 ST 段的斜率：1 为上坡，2 为平坦，3 为下坡；是否患有心脏病：1 为患有心脏病，2 为未患有心脏病。

惩罚函数参数 k，采用设定网格进行搜索，以调控其惩罚强度。然后，为确保模型稳健性，采用十折交叉验证的方式在训练集中评估各参数组合的性能，并选择最优组合。最后，利用估计的调谐参数计算分类超平面。数值实验评价指标见表 2。

Table 2. Evaluation metrics of three classification models
表 2. 三种分类模型评价指标

模型	测试误差(%)	灵敏度(%)	特异性(%)	AUC 值
SVM-L_2-L_1	14.11	84.38	86.21	0.9052
SVM-L_2-SCAD	15.39	84.38	86.21	0.8782
SVM-L_2-modSCAD	**11.48**	**90.62**	**88.72**	**0.9246**

结果分析。由数值实验结果可以看出，SVM-L_2-modSCAD 模型在测试误差(11.48%)、灵敏度(90.62%)、特异性(88.72%)以及 AUC 值(92.46%)表现最优，表明该模型在整体分类准确率以及识别心脏病患者(正类)和非患病个体(负类)方面更为准确，同时，其 AUC 值也最高(0.9246)。此外，SVM-L_2-modSCAD 模型成功筛选出 7 个关键特征，包括性别、胸痛类型、运动诱发心绞痛、ST 段压低、ST 段斜率、主要血管数和地中海贫血类型，这些关键特征在临床上均被认为与心脏病密切相关。其中，ST 段斜率是评估心肌缺血和心肌功能异常的重要指标，在心脏病早期筛查中具有重要意义。值得注意的是，该特征未被 SVM-L_2-L_1 与 SVM-L_2-SCAD 模型识别出来，这说明 SVM-L_2-modSCAD 模型在特征选择上具备更强的识别能力和解释力。SVM-L_2-modSCAD 模型通过合理控制惩罚项的非凸程度，不仅在维持良好预测性能的同时有效减少了冗余变量，还增强了模型在临床诊断中的实用性和可解释性，显示出其在医疗数据建模中的综合优势。此外，从复杂度角度来看，本文所提 SVM-L_2-modSCAD 模型(线性核)与上述已有模型(线性核)的时间复杂度均为 O (np) (n 为样本数，p 为特征数)，空间复杂度也未变化。从运行时间来看，SVM-L_2-L_1 模型运行时间为 2.31 秒，SVM-L_2-SCAD 模型与 SVM-L_2-modSCAD 模型运行时间均为 10.02 秒，本文所提 SVM-L_2-modSCAD 模型与 SVM-L_2-SCAD 模型运行时间相同，比 SVM-L_2-L_1 模型运行时间长，主要是因为模型中加入了信息准则(AIC 和 BIC)选择最优调谐参数 λ_2 时的循环判断，增加了模型运行时间。

综上所述，SVM-L_2-modSCAD 模型在复杂度不变、运行时间差别不大的情况下，具备较强特征筛选

能力，在兼顾模型拟合与泛化性能的同时实现了更优的分类效果，体现出更好的实用性。三种分类模型和算法特征选择数目与各参数具体结果如下表 3。

Table 3. Parameter values of the three models
表 3. 三种模型的参数值

模型	参数 λ_1	调谐参数 λ_2	特征数量	k
SVM-L_2-L_1	7.5646	-	6	-
SVM-L_2-SCAD	0.8111	0.0042	6	-
SVM-L_2-modSCAD	0.8111	0.4832	7	1.5

5. 结论

本文提出了基于岭惩罚与修正的 SCAD 惩罚的软间隔 SVM 分类模型与算法，该分类模型与算法同时考虑了样本信息与先验信息，发挥了修正的 SCAD 惩罚在减少估计偏差和保持稀疏性方面的优势，同时利用岭惩罚提升了算法的在特征选择上的稳定性，是基于岭惩罚与传统的 SCAD 惩罚的软间隔 SVM 的重要拓广。通过心脏病诊断实例验证，相对于 SVM-L_2-L_1 与 SVM-L_2-SCAD 分类算法，该分类算法具有更优的灵敏度、特异度和分类能力。尽管如此，本文研究尚存在局限性，如，本文所选择的数据规模相对较小，在大规模数据场景中，尤其是涉及高维特征或超参数网格密集搜索时，模型训练成本会急剧上升，进而影响本文所提方法的可扩展性。

基金项目

国家重点研发计划项目(项目编号：2023YFC3209403-04-05)。

参考文献

[1] Tikhonov, A. and Arsenin, V. (1977) Solutions of Ill-Posed Problems. Wiley.

[2] Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer. https://doi.org/10.1007/978-1-4757-2440-0

[3] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/BF00994018

[4] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[5] Becker, N., Toedt, G., Lichter, P. and Benner, A. (2011) Elastic SCAD as a Novel Penalization Method for SVM Classification Tasks in High-Dimensional Data. *BMC Bioinformatics*, **12**, Article No. 138. https://doi.org/10.1186/1471-2105-12-138

[6] Ng, C.T. and Yu, C.W. (2014) Modified SCAD Penalty for Constrained Variable Selection Problems. *Statistical Methodology*, **21**, 109-134. https://doi.org/10.1016/j.stamet.2014.05.001

[7] Zhang, H.H., Ahn, J., Lin, X. and Park, C. (2006) Gene Selection Using Support Vector Machines with Non-Convex Penalty. *Bioinformatics*, **22**, 88-95. https://doi.org/10.1093/bioinformatics/bti736

[8] Mangasarian, O.L. and Musicant, D.R. (2001) Lagrangian Support Vector Machines. *Journal of Machine Learning Research*, **1**, 161-177. https://dl.acm.org/doi/10.1162/15324430152748218

[9] 梅瑞婷, 徐扬, 王国长. 基于LASSO-SVM模型的银行定期存款电话营销预测[J]. 统计学与应用, 2016, 5(3): 289-298.

[10] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., *et al*. (1989) International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease. *The American Journal of Cardiology*, **64**, 304-310. https://doi.org/10.1016/0002-9149(89)90524-9