p 范数正则化支持向量机分类算法

刘建伟 [1] 李双成 [1] 罗雄麟 [1]

摘 要 L_2 范数罚支持向量机 (Support vector machine, SVM) 是目前使用最广泛的分类器算法之一, 同时实现特征选择和分类器构造的 L_1 范数和 L_0 范数罚 SVM 算法也已经提出. 但是, 这两个方法中, 正则化阶次都是事先给定, 预设 $p = 2$ 或 $p = 1$. 而我们的实验研究显示, 对于不同的数据, 使用不同的正则化阶次, 可以改进分类算法的预测准确率. 本文提出 p 范数正则化 SVM 分类器算法设计新模式, 正则化范数的阶次 p 可取范围为 $0 < p \leq 2$. 使用网格法选择模型参数值, 使用迭代再权方法求解分类器目标函数, 找出最小分类预测误差的模型参数值. 在实际数据集上的实验结果验证了提出算法能够同时实现分类预测和特征选择, 性能优于 L_2 范数罚 SVM, L_1 范数罚 SVM 和 L_0 范数罚 SVM.

关键词 迭代再权方法, p 范数 ($0 < p \leq 2$), 支持向量机, 特征选择, 稀疏化模型, 高维小样本数据

DOI 10.3724/SP.J.1004.2012.00076

Classification Algorithm of Support Vector Machine via p-norm Regularization

LIU Jian-Wei[1] LI Shuang-Cheng[1] LUO Xiong-Lin[1]

Abstract The L_2 penalty support vector machine (SVM) algorithm is one of the most widely used learning algorithms, meanwhile L_1 norm and L_0 norm penalty support vector machines have been devised, which achieve simultaneously feature selection and classifier construction. However, in both methods, the regularization parameter is predetermined, i.e., the default $p = 2$ or $p = 1$. Our experimental study shows that different data, using a different regularization of order, can improve prediction accuracy of the classification algorithm. In this paper, new classifier design pattern of SVM based on p-norm regularization is proposed, where $0 < p \leq 2$. We design grid method to select parameter values of model, use the iterative reweighted method to solve classification object function then discover the right parameter values of model at the minimum prediction error. The performance of classification and feature selection on real datasets indicate that the devised algorithm is better than L_2-norm, L_1-norm, and L_0-norm SVM.

Key words Iterative reweighted method, p-norm ($0 < p \leq 2$), support vector machine (SVM), feature selection, sparse model, high-dimensional small sample dataset

Vapnik 等提出的支持向量机 (Support vector machine, SVM) 为当前主流的机器学习算法 (以下简称为 L_2-SVM)[1−2]. 与此同时, 基于 L_0 或者 L_1 范数正则化的同时实现特征选择和分类以及回归的算法 (以下简称为 L_1-SVM、L_0-SVM) 也相继被提出[3−6]. Liu 等提出了组合 L_0、L_1 范数组成复合罚函数的 SVM 分类和回归算法[7], 同时提出根据数据特性选择 L_1、L_2 范数的 SVM 分类算法[8], 该算法与我们提出的算法不同, 正则化阶次只能取 $p = 2$ 或者 $p = 1$, 而我们提出的算法 p 的取值范围为 $0 < p \leq 2$. 另一个值得关注的研究热点是引入非凸罚函数构造分类回归算法的研究[9−11], 在这些算法中罚函数为分数范数 L_p, $0 < p < 1$.

所有以上文献均对某一种范数正则化阶次分类算法进行研究, 但根据我们的实验研究, 对于不同的数据, SVM 的分类准确率随正则化阶次不同而不同, 范数 L_p 的阶次可能的范围 $0 < p \leq 2$, 如何能够在不同的数据集上, 求解分类准确率最优的正则化范数的阶次, 并实现 $0 < p \leq 2$ 不同范数正则化的分类算法, 仍然未解决.

关于迭代再权最小二乘的研究, Holland 等于 1977 年提出迭代再权最小二乘回归算法[12], 之后, Gorodnitsky 等提出仿射比例再权稀疏信号重构算法[13], 该算法可应用于 $p = 0$, $0 < p < 2$ 的各种范数, 且有很好的收敛性. Daubechies 等相继提出了基于迭代再权的压缩传感算法[14−17]. 为解决问题

$$\min_x ||\boldsymbol{x}||_p \ , \ \text{s. t. } \boldsymbol{y} = \Phi \boldsymbol{x} \tag{1}$$

Chartrand 提出迭代再权算法的迭代公式为 $\boldsymbol{x}^{k+1} := \arg\min_{\boldsymbol{x}} ||\boldsymbol{y} - \Phi\boldsymbol{x}||_2^2 + \lambda \sum_i w_i^{(k)} x_i^2$, 权 $w_i^{(k)}$ 定义为 $w_i^{(k)} := \left[(x_i^{(k+1)})^2 + \varepsilon_{(k+1)}\right]^{-1}$, Daubechies

收稿日期 2010-12-24 录用日期 2011-08-30
Received December 24, 2010; accepted August 30, 2011
国家自然科学基金 (21006127, 20976193), 中国石油大学 (北京) 基础学科研究基金项目资助
Supported by National Natural Science Foundation of China (21006127, 20976193), Basic Scientific Research Foundation of China University of Petroleum
本文责任编委 乔红
Recommended by Associate Editor QIAO Hong
1. 中国石油大学自动化研究所 北京 102249
1. Research Institute of Automation, China University of Petroleum, Beijing 102249

定义权为 $w_i^{(k)} := \left[(x_i^{(k+1)})^2+\varepsilon_{(k+1)}^2\right]^{-1/2}$, Candes 提出的算法权定义为 $w_i^{(k+1)} := \left[\left|(x_i^{(k+1)})\right|+\varepsilon_{(k+1)}\right]^{-1}$. 当前, 如何从分子生物学实验数据发现基因表达、基因调控规律是目前生物信息学面临的巨大挑战, 而强有力的机器学习理论恰恰有了用武之地, 如运用 SVM 对癌症分类的文献有 [18−19]. 另外, 人脸识别数据具有高维和稀疏等特点, 如何实现特征选择、改进识别的准确性也是一个富于挑战性的研究内容[20−22].

针对现有的 SVM 分类算法中常把正则化阶次 p 事先定为 0, 1 或 2, 即 0 范数 SVM, 1 范数 SVM 或 2 范数 SVM, 通过大量实验, 我们发现: 最优的分类效果并非出现在 $p=1$ 或 2 处, 而是可能出现在 (0, 2] 的任意位置, 具体出现在什么位置, 则需要在 p 的取值范围内求解 SVM 目标函数, 寻找最优分类器性能时的 p 值.

本文的主要工作如下:

1) 提出了 p 范数正则化 SVM 分类器算法设计模式, 分类算法能够在 $0<p\le 2$ 范围内, 选择使得分类器预测误差最小的正则化范数的阶次 p;

2) 提出了求解 p 范数正则化 SVM 目标函数的迭代再权算法, 对 SVM 分类算法中的两个可调参数 — 损失函数相对于模型复杂性之间的权衡参数 c 和正则化阶次 p, 提出了关于这两个参数的取值范围内的寻优算法;

3) 由于分类器误差最小时的正则化阶次为分数, 因而训练得到的模型向量 $\boldsymbol{w}$ 的大多数分量为零, 这样使得模型向量 $\boldsymbol{w}$ 分量为零的维所对应的样本的相应坐标分量不再参与模型的构造, 只有那些模型向量 $\boldsymbol{w}$ 的分量不为零的维所对应的样本的坐标分量才参与模型构造, 故实现了特征选择;

4) 把提出的分类算法应用于有代表性的三组数据上: Prostate-Tumor 数据集、YaleB0102 人脸数据集和 YaleB2627 人脸数据集, 实验结果验证了提出算法能够同时实现分类预测和特征选择, 而最小预测错误率不是在某个固定的正则化阶次 p 上;

5) 选择了四组基因序列癌症预测数据: 乳腺癌数据、黑素瘤数据、淋巴癌数据和结肠癌数据, 把提出的分类算法与标准的 SVM, L_1 范数 SVM, L_0 范数 SVM 进行了实验比较, 比较了预测错误率和特征选择错误率, 实验结果表明了提出算法的优越性.

1　L_2-SVM, L_1-SVM 和 L_0-SVM

假设样本集为 $S=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_t,y_t),\cdots\}\subseteq\mathbf{R}^n\times\{\pm1\}$, $\boldsymbol{x}_i$ 为样例, y_i 为对应的类标签. L_2-SVM 求解的问题为

$$\min\|\boldsymbol{w}\|_2^2+C\xi_i^q$$
$$\text{s. t. } y_i\cdot(\langle\boldsymbol{x}_i,\boldsymbol{w}\rangle+b)\ge 1-\xi_i\quad i=1,\cdots,m \tag{2}$$

其中, $q=1$ 时为一次软间隔, $q=2$ 时为二次软间隔, 本文只考虑二次软间隔情形, 其他两种情况的讨论与之类似. 求解约束目标函数得到最优分类器模型参数: 权向量 $\boldsymbol{w}$ 和偏置 b. L_2-SVM 的研究主要是基于分解理论对二次优化算法本身的改进和算法的实际应用, 基于分解理论构造的算法典型的有: 块算法 (Chunking algorithm)[23], 固定工作样本集算法[24], SMO[25], SVMperf[26], Pegasos[27], DCDL2, DCDL2[28] 和 BMRM[29].

L_1-SVM 求解的问题为

$$\min\sum_{i=1}^{m}[1-y_i\cdot(\langle\boldsymbol{x}_i,\boldsymbol{w}\rangle+b)]$$
$$\text{s. t. }\ \|\boldsymbol{w}\|_1\le s \tag{3}$$

对 L_1-SVM 的研究主要集中于两类[30−31] 和多类[32−33] 算法的设计. L_1-SVM 算法的主要思想是利用当 $0\le s\le\infty$ 时, 随着 s 改变, 对于 Γ_1, $\boldsymbol{w}$ 是 s 的分段线性函数的特点, 求解整个解路径, 用于特征选择时, L_1-SVM 没有使用非线性核映射.

L_0-SVM 求解的问题为

$$\min\|\boldsymbol{w}\|_0+C\xi_i^q$$
$$\text{s. t. } y_i\cdot(\langle\boldsymbol{x}_i,\boldsymbol{w}\rangle+b)\ge 1-\xi_i,\ \ i=1,\cdots,m \tag{4}$$

L_0-SVM 问题的求解为非多项式可解 (NP) 问题[34], 解决的办法是采用近似放松的方法, 使得 L_0-SVM 问题的求解过程转化为一系列 L_2-SVM 求解过程[35], 在文献 [3] 中, 用于特征选择时, L_0-SVM 也没有使用高斯核非线性映射.

2　p 范数正则化 SVM 分类算法

2.1　p 范数的导数求解

p 范数定义为

$$h(\boldsymbol{w})=\|\boldsymbol{w}\|_p=\left[\sum_{j=1}^{n}|w_j|^p\right]^{\frac{1}{p}} \tag{5}$$

令 $k(\boldsymbol{w})=h^p(\boldsymbol{w})$, 当 $0<p\le 2$ 时, 其导数为

$$\frac{\partial k(\boldsymbol{w})}{\partial w_j}=p\,|w_j|^{p-1}\cdot\operatorname{sgn}(w_j) \tag{6}$$

由于

$$\mathrm{sgn}(w_j)=\frac{w_j}{|w_j|} \tag{7}$$

所以

$$\frac{\partial k(\boldsymbol{w})}{\partial w_j}=p\,|w_j|^{p-2}\cdot w_j \tag{8}$$

2.2 p 范数正则化迭代再权 SVM 分类算法

我们提出的 p 范数正则化迭代再权 SVM 分类算法使用线性核, 数据不经过非线性核映射. 理由是: 首先核的选择问题是一个难题, 不同的数据可能使用不同的核, 而且即使核的形式确定后, 核参数的选择又是一个难题; 其次, 要实现特征选择, 必须显式求出 $\boldsymbol{w}$, 这对于高斯核等无穷维核映射, 现在还无法实现[36].

假定 $\boldsymbol{w},\boldsymbol{x}_i\in\mathbf{R}^n$, $y_i\in\{\pm1\}$, 样本线性不可分, 我们提出的 p 范数正则化迭代再权 SVM 分类算法的目标函数和约束条件为

$$\min\frac{1}{2}\|\boldsymbol{w}\|_p^p+\frac{C}{2}\sum_{i=1}^{m}\xi_i^2$$

$$\text{s. t. }\ 1-y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b)-\xi_i\le 0,\quad 0\le i\le m \tag{9}$$

其中, $0<p\le 2$, 注意到上述目标函数中, 当取 $p=1$ 时, 此问题变为标准的 L_1 范数罚 SVM. 当取 $p=2$ 时, 此问题变为标准的 L_2 范数罚 SVM, 此问题在 $0<p<1$ 时为非凸问题, $p=1$ 时为非平滑问题, $p=2$ 时为二次规划问题, 那么是否存在有效的方法求解该问题? 我们提出的 p 范数正则化迭代再权 SVM 分类算法由一系列迭代过程组成, 算法每一次迭代总是求解一个标准的二范数 SVM 二次对偶问题, 一旦解出 $\boldsymbol{w}_{t-1}$, 下一次求解的目标函数中权值 $|\boldsymbol{w}_t|^{p-2}$ 用近似值 $|\boldsymbol{w}_{t-1}|^{p-2}$ 代替, 而在初次迭代中, 算法任意置一个初值 $\boldsymbol{w}_0$ 或由标准的二范数 SVM 解作为 $\boldsymbol{w}$ 的初值.

为了避免繁琐的符号表示, 以下讨论中, 目标函数中的各变量 $\boldsymbol{w}$, b, $\boldsymbol{\xi}$ 并未用标号表示出迭代次数的序号 t.

第 t 次迭代, 上述最优化问题的无约束拉格朗日函数为

$$\begin{aligned}L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=&\frac{1}{2}\sum_{j=1}^{n}|w_j|^p+\frac{C}{2}\sum_{i=1}^{m}\xi_i^2-\\&\sum_{i=1}^{m}\beta_i\xi_i+\sum_{i=1}^{m}\alpha_i\left[1-y_i\left(x_i^{\mathrm{T}}\boldsymbol{w}-b\right)-\xi_i\right]+\\&\sum_{i=1}^{m}\mu_i\left(\xi_i-1\right)=\\&\frac{1}{2}\sum_{j=1}^{n}|w_j|^{p-2}w_j^2+\frac{C}{2}\sum_{i=1}^{m}\xi_i^2-\\&\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+\sum_{i=1}^{m}\alpha_i y_i b+\sum_{i=1}^{m}\alpha_i-\\&\sum_{i=1}^{m}\alpha_i\xi_i-\sum_{i=1}^{m}\beta_i\xi_i+\sum_{i=1}^{m}\mu_i\xi_i-\sum_{i=1}^{m}\mu_i\end{aligned} \tag{10}$$

把 $L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})$ 作为第 t 次迭代的目标函数, 把 $L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})$ 中的 $|w_j|^{p-2}$ 用次目标函数的解 $|w_{t-1,j}|^{p-2}$ 近似替代, 令 $|\hat{w}_j|^{p-2}$ 代替 $|w_{t-1,j}|^{p-2}$, 以便能够更好地表示提出算法的含义, 并引入向量表示, 得:

$$\begin{aligned}&L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=\\&\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\begin{bmatrix}|\hat{w}_1|^{p-2}&&\\&\ddots&\\&&|\hat{w}_n|^{p-2}\end{bmatrix}\boldsymbol{w}-\\&\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}X\boldsymbol{w}+\frac{C}{2}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\xi}-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\xi}+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\xi}+\\&\sum_{i=1}^{m}\alpha_i+\sum_{i=1}^{m}\alpha_i y_i b-\sum_{i=1}^{m}\mu_i=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}V\boldsymbol{w}-\\&\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}X\boldsymbol{w}+\frac{C}{2}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\xi}-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\xi}+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\xi}+\\&\sum_{i=1}^{m}\alpha_i+\sum_{i=1}^{m}\alpha_i y_i b-\sum_{i=1}^{m}\mu_i\end{aligned} \tag{11}$$

其中, $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_m]^{\mathrm{T}}$, $X=[x_1^{\mathrm{T}},\cdots,x_m^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{\xi}=[\xi_1,\cdots,\xi_m]^{\mathrm{T}}$, $\boldsymbol{\beta}=[\beta_1,\cdots,\beta_m]^{\mathrm{T}}$, $V=\begin{bmatrix}|\hat{w}_1|^{p-2}&&\\&\ddots&\\&&|\hat{w}_n|^{p-2}\end{bmatrix}$.

对拉普拉斯函数就变量 $\boldsymbol{w}$, b, $\boldsymbol{\xi}$ 求导:

$$\frac{\partial L}{\partial\boldsymbol{w}}=\boldsymbol{w}^{\mathrm{T}}V-\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}X \tag{12}$$

$$\frac{\partial L}{\partial b}=\sum_{i=1}^{m}\alpha_i y_i \tag{13}$$

$$\frac{\partial L}{\partial\boldsymbol{\xi}}=C\boldsymbol{\xi}^{\mathrm{T}}-\boldsymbol{\alpha}^{\mathrm{T}}-\boldsymbol{\beta}^{\mathrm{T}}+\boldsymbol{\mu}^{\mathrm{T}} \tag{14}$$

利用 KKT 条件, 得:

$$\frac{\partial L}{\partial\boldsymbol{w}}=0\Rightarrow\boldsymbol{w}^{\mathrm{T}}=\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}XV^{-1} \tag{15}$$

$$\frac{\partial L}{\partial b}=0\Rightarrow\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha}=0 \tag{16}$$

$$\frac{\partial L}{\partial \boldsymbol{\xi}}=0\Rightarrow \boldsymbol{\xi}=\frac{1}{C}(\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{\mu}) \tag{17}$$

把上述条件代入 $L(\boldsymbol{w},b,\boldsymbol{\xi})$, 消去 $\boldsymbol{w}$, b, $\boldsymbol{\xi}$, 得对偶问题:

$$\begin{aligned}
&\max\ L_D(\boldsymbol{\alpha},\boldsymbol{\beta})=-\frac{1}{2}\mathrm{diag}\{\boldsymbol{y}\}XV^{-1}X^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}\boldsymbol{\alpha}-\\
&\frac{1}{2C}(\boldsymbol{\alpha}^{\mathrm{T}}+\boldsymbol{\beta}^{\mathrm{T}}-\boldsymbol{\mu}^{\mathrm{T}})(\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{\mu})+\sum_{i=1}^{m}\alpha_i-\\
&\sum_{i=1}^{m}\mu_i=-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}XV^{-1}X^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}\boldsymbol{\alpha}-\\
&\frac{1}{2C}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}-\frac{1}{2C}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}-\frac{1}{2C}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}+\sum_{i=1}^{m}\alpha_i-\\
&\sum_{i=1}^{m}\mu_i-\frac{1}{2C}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\beta}+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\alpha}-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu}-\\
&\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\alpha}-\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta})=-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}XV^{-1}X^{\mathrm{T}}\times\\
&\mathrm{diag}\{\boldsymbol{y}\}\boldsymbol{\alpha}-\frac{1}{2C}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}-\frac{1}{2C}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}-\\
&\frac{1}{2C}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}-\frac{1}{2C}\left[\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\mu}^{\mathrm{T}}\right]\times\\
&\begin{bmatrix}0^{m\times m}&1^{m\times m}&-1^{m\times m}\\1^{m\times m}&0^{m\times m}&-1^{m\times m}\\-1^{m\times m}&-1^{m\times m}&0^{m\times m}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\\\boldsymbol{\mu}\end{bmatrix}+\\
&\underbrace{[1,1,\cdots]}_{n}\boldsymbol{\alpha}-\underbrace{[1,1,\cdots]}_{n}\boldsymbol{\mu}
\end{aligned}\tag{18}$$

其中, $0^{m\times m}$ 为 $m\times m$ 的全零元素矩阵, $1^{m\times m}$ 为 $m\times m$ 的全 1 元素矩阵. 约束条件为

$$\begin{cases}0\le\alpha_i,\\0\le\beta_i,\\0\le\mu_i,\\\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha}=0,\end{cases}\quad i=1,2,\cdots,m \tag{19}$$

由于 $\boldsymbol{\xi}=\frac{1}{C}(\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{\mu})$, 且 $0\le\alpha_i$, $0\le\beta_i$, $0\le\mu_i$, $0\le\xi_i\le 1$, 故可推得:

$$\begin{cases}0\le\alpha_i\\0\le\beta_i\\0\le\mu_i\\\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha}=0\end{cases}\Leftrightarrow\begin{cases}0\le\alpha_i\le C,\\0\le\beta_i\le C,\\0\le\mu_i\le C,\\\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha}=0,\end{cases}\quad i=1,2,\cdots,m \tag{20}$$

$L_D(\boldsymbol{\alpha},\boldsymbol{\beta})$ 的两边同时乘以 -1, 把上述最大问题转化为最小问题, 得最终的对偶二次目标函数为

$$\begin{aligned}
&\min L_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}XV^{-1}X^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}\boldsymbol{\alpha}-\\
&\frac{1}{2C}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}-\frac{1}{2C}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}-\frac{1}{2C}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}-\frac{1}{2C}\times\\
&\left[\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\mu}^{\mathrm{T}}\right]\begin{bmatrix}0^{m\times m}&1^{m\times m}&-1^{m\times m}\\1^{m\times m}&0^{m\times m}&-1^{m\times m}\\-1^{m\times m}&-1^{m\times m}&0^{m\times m}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\\\boldsymbol{\mu}\end{bmatrix}+\\
&\underbrace{[1,1,\cdots]}_{n}\boldsymbol{\alpha}-\underbrace{[1,1,\cdots]}_{n}\boldsymbol{\mu}
\end{aligned}\tag{21}$$

约束条件为

$$\begin{cases}0\le\alpha_i\le C,\\0\le\beta_i\le C,\\0\le\mu_i\le C,\\\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha}=0,\end{cases}\quad i=1,2,\cdots,m \tag{22}$$

采用 JohnPlatt 的序列最小最优化算法求解二次 Wolfe 对偶问题. 具体的最优化算法见算法 1. 给定初始值 $(\boldsymbol{w}^{(0)},b^{(0)})$, 分类算法迭代选择两个最不符合 KKT 条件的训练样本作为工作集 S, 由 $(\boldsymbol{w}^{(t)},b^{(t)})$ 迭代再权求解 $L_D(\boldsymbol{\alpha},\boldsymbol{\beta})$, 得到 $(\boldsymbol{w}^{(t+1)},b^{(t+1)})$, 直到满足算法预设的停止条件, 从而得到最终的模型参数 $\boldsymbol{w}$ 和 b.

算法 1. 迭代再权 p 范数正则化 SVM 算法

输入参数. 样本 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}\subseteq\mathbf{R}^n\times\{\pm1\}$, 停止标准 ς;

输出. 模型参数向量 $\boldsymbol{w}$ 和 $\boldsymbol{b}$;

对 $\boldsymbol{w}$ 置初值: 令 $w_j^{(0)}=1/\sqrt{n}$, $j=1,\cdots,n$, 或者使用标准的 L_2 范数罚 SVM 解作为 $\boldsymbol{w}^{(0)}$; 重复利用 $\boldsymbol{w}^{(t)}$, 使用 John Platt 提出的序列最小最优化算法求解 $\boldsymbol{\alpha}^{(t+1)}$, $\boldsymbol{\beta}^{(t+1)}$, $\boldsymbol{\mu}^{(t+1)}$, 直到 $\left\|\boldsymbol{w}^{(t-1)}-\boldsymbol{w}^{(t)}\right\|<\varsigma$; 计算 $\boldsymbol{w}^{\mathrm{T}}=\boldsymbol{a}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}XV^{-1}$; 计算 $b=\frac{1}{|S|}\sum_{k\in S}(\boldsymbol{x}_k\mathrm{T}\boldsymbol{w}-y_k+y_k\cdot\max\left\{0,\frac{\alpha_k+\beta_k-\mu_k}{C}\right\})$.

算法中要构造 $n\times n$ 维的对角矩阵并参与矩阵运算, n 为样本的特征数, 如前列腺癌 (Prostate-Tumor) 数据集的 $n=10\,509$, 如此大的内存资源消耗, 用 Matlab 以及其他软件无法在个人 PC 机上运行. 算法实现上考虑采用分块技术, 实现对超高维矩阵的处理. 程序中, 块大小可以由算法的使用者自行决定. 如前列腺癌数据集的块大小定为 2 000, 数据被分成 6 块, 经过块划分, 各块包含的特征数分别为 2 000, 2 000, 2 000, 2 000, 2 000, 509.

$L_D(\boldsymbol{\alpha},\boldsymbol{\beta})$ 中需要分块计算的是其第一项, 定义 $L_D(\boldsymbol{\alpha},\boldsymbol{\beta})$ 中的第一项为矩阵 O:

$$O=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}XV^{-1}X^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{y}\}\boldsymbol{\alpha} \tag{23}$$

令

$$M=\mathrm{diag}\{\boldsymbol{y}\}X \tag{24}$$

则矩阵 O 可表示为

$$O=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}MVM^{\mathrm{T}}\boldsymbol{\alpha} \tag{25}$$

其中, V 为 $n\times n$ 对角矩阵, M 为 $m\times n$ 矩阵. 算法设计时, 把矩阵 MVM^{T} 分为 r 个块, 定义 MVM^{T} 的第 i 个数据块 $R(i)$:

$$R(i)=M(i)V(i)M(i)^{\mathrm{T}} \tag{26}$$

其中, 矩阵 $M(i)$ 为矩阵 M 的子矩阵, 该子矩阵由 M 的 $2\,000(i-1)+1$ 列至 $2\,000i$ 列组成; 矩阵 $V(i)$ 为矩阵 V 的子矩阵, 该子矩阵由 V 的 $2\,000(i-1)+1$ 行至 $2\,000i$ 行组成. 用公式 $R(i)=M(i)V(i)M(i)^{\mathrm{T}}$ 计算全部数据块 i, $i=1,\cdots,m$, 最终 $MVM^{\mathrm{T}}=R(1)+R(2)+\cdots+R(r)$, 最后即可计算出矩阵 O.

采用此方法, 内存消耗减少到不采用分块算法所需内存的 $1/r^2$, 算法所需计算量和耗时也减少了. 原始矩阵 V 中有 n^2-n 个零元素, 矩阵直接相乘, 矩阵中的零元素也将参与计算; 分块后, 矩阵 V 中, $V(i)$ $(i=1,\cdots,r)$ 以外的元素全部为零元素, 这些零元素不参与计算, 算法复杂度减少到不分块时算法复杂度的 $1/r^2$.

2.3 模型参数选择过程

文献 [36] 给出把 m 个样本组成的样本集分为个数为 m_1 的训练集和个数为 m_2 的交叉校验集, 在不同的给定可调参数值下, 在训练数据集上用逻辑斯蒂分类方法训练得到模型向量, 而后在所有模型向量中选择具有最低校验误差的可调参数值作为测试过程的可调参数值. 我们提出的 L_p 正则化 SVM 也使用这种模式确定可调参数的值.

p 范数正则化迭代再权 SVM 分类算法共有两个可调参数: 损失函数相对于模型复杂性之间的权衡参数 c 和正则化阶次 p, 由于 p 范数正则化迭代再权 SVM 分类算法同时实现分类和特征选择, 分类算法实现功能中有两个性能指标: 预测误差 $err(c,p)$ 和特征选择个数 $feat(c,p)$, 它们均是 c 和 p 的函数. 分类算法需要根据性能指标选取适当的 c 和 p 值, 实现最优的分类误差和最优的特征选择. 解路径追踪算法使用网格法, 把 c 和 p 的取值范围人为划分, 在特定的 c, p 和训练样本上调用算法 1 求解 $\boldsymbol{w}$ 和 b, 然后用 $\boldsymbol{w}$ 和 b 在测试样本上预测样例的类标签, 最终使得 $err(c,p)$ 最小. 分类算法中, c 共取 15 个可能的固定值, 以等比增长分布, 初始值为 0.0625, 终值为 1 024, 公比为 2. 对于每一个固定的 c 值选取不同的 p 值, p 值以等差增长分布, 初始值为 0.2, 终值为 2, 公差为 0.2, 各 p 值点间用直线连接, 构成解路径 $err(c,p)$ 和 $feat(c,p)$, 全部解路径由 150 个节点组成, 在每个节点上, 做 60 次重复实验, 将平均值作为该节点的最终结果. 获得全部节点数据后, 绘制解路径, 找出解的变化规律和最优解.

c 值的选取范围的确定出于以下考虑: 首先, c 值决定 $\|\boldsymbol{w}\|_p^p$ 与 $\sum_{i=1}^m\xi_i^2$ 的比例关系, 所以 c 按照等比关系取值, 更能说明分类效果随 $\|\boldsymbol{w}\|_p^p/\sum_{i=1}^m\xi_i^2$ 的变化而变化; 其次, 大量实验显示, 当 $c>8$ 时, 算法的分类效果不再随 c 值的增大而发生明显变化, 当 $0<c<8$ 时, 定义 ph 为区间 $(0,2)$ 之间的某个值, 有两种情况: 1) $p<ph$ 时, 算法的分类效果随 c 值的增大而变好; 2) $p>ph$ 时, 算法的分类效果不随 c 值的变化而发生明显变化. 因此, 最优解对应的 c 值必然出现在 0.0625 与 1 024 之间, 因此, 将 c 值确定为上述 15 个可能值.

p 的选值范围的确定出于以下考虑: 本文致力于训练稀疏模型, 从而达到主特征筛选, 减小预测错误率的目的. 大量实验证明, $p>2$, 全部特征都被选作主特征, 获得的模型不稀疏. 所以将 p 的取值范围定在区间 $(0,2]$.

2.4 算法复杂性分析

标准的 SMO 算法计算复杂性为 $\mathrm{O}(m\times\bar{L})$, 其中, m 为样本个数, $\bar{L}$ 为 SMO 算法迭代过程中支持向量的个数. 实际程序运行表明, 只需要几次到十余次迭代过程就能基本收敛. 由于 SMO 算法也要交叉校验选取 c 值, 总体上, 我们的 L_p-SVM 算法的 SMO 算法子回路部分的复杂性为 $\mathrm{O}(m\times\bar{L})$. 计算 $\boldsymbol{w}$ 复杂性为 $\mathrm{O}(m^2n+n^3)$, 总的算法复杂性为 $\mathrm{O}(m^2n+n^3+m\times\bar{L})$, 实际上由于支持向量才参与运算, 且样例的很多维随着算法的计算过程变得稀疏, 故给出的复杂性为最坏情况的算法复杂性.

2.5 L_p 范数正则化 SVM 算法抽样复杂性界的理论分析

我们的理论分析基于覆盖数理论[37−38], 为了方便讨论, 先给出相关的定义和引理.

定义 1. 令 M 为度量为 d 的度量空间, 给定样例集 $X=\lfloor\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m\rfloor\in D$, F 为 D 上函数类, F 的值域为 $[-M,M]\in\mathbf{R}$, $f(\boldsymbol{w},X)=\lfloor f(\boldsymbol{w},\boldsymbol{x}_1),\cdots,f(\boldsymbol{w},\boldsymbol{x}_m)\rfloor$ p 范数覆盖数 $N_p(f,\epsilon,X)$ 为满足对于 $\forall\boldsymbol{w}$, $\exists\boldsymbol{v}_i$, 且使

$$\|d\left(f(\boldsymbol{w},X),\boldsymbol{v}_i\right)\|_p=\left[\sum_{k=1}^n d\left(f(\boldsymbol{w},\boldsymbol{x}_k),v_{i,k}\right)^p\right]^{\frac{1}{p}}\le n^{\frac{1}{p}}\varepsilon \tag{27}$$

成立的最小向量集 $V=[\boldsymbol{v}_1,\cdots,\boldsymbol{v}_p]$.

定义 2. 假定样例集 $X=\lfloor\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m\rfloor$, 由某

个固定、未知的概率分布为 D 的样例集中抽取得到, 相应的类标签集为 $Y=\lfloor \boldsymbol{y}_1,\cdots,\boldsymbol{y}_m \rfloor$. 定义铰链损失为 $L=(\boldsymbol{x}_i,y_i,f(\boldsymbol{x}_i))=\max(0,1-yf(\boldsymbol{x}_i))$, 则期望风险定义为

$$R\left(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i)\right)=\int_{X,Y} L\left(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i)\right)\mathrm{d}P(X,Y) \tag{28}$$

经验风险定义为

$$\hat{R}\left(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i)\right)=\frac{1}{n}\sum_{i=1}^{m} L\left(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i)\right) \tag{29}$$

引理 1[39]. 对于 $\forall \boldsymbol{w}\in\mathbf{R}^n$, $0<p<q<2$, 则有

$$\|\boldsymbol{w}\|_p \geq \|\boldsymbol{w}\|_q \tag{30}$$

引理 2[40]. 对于 $\forall\,\varepsilon$, 概率分布 D, 下式成立

$$P\left[\exists f\in F:\left|\hat{R}(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i))-R\left(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i)\right)\right|>\varepsilon\right]\leq 8\mathrm{E}\left[N_1(F,\frac{\varepsilon}{8L},X)\right]\exp\left[-\frac{m\varepsilon^2}{512M^2}\right] \tag{31}$$

这里 $M=\sup_{\boldsymbol{w},X}F(\boldsymbol{w},x)-\inf_{\boldsymbol{w},X}F(\boldsymbol{w},x)$, L 为风险中的损失函数的利普希茨常数.

引理 3[41]. 定义函数类

$$\Lambda=\left\{\lambda:\lambda(x)=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x},\boldsymbol{x}\in\mathbf{R}^n\right\}$$

且 $\|\boldsymbol{x}\|_p\leq c$, $\|\boldsymbol{\theta}\|_q\leq d$, $1/p+1/q=1$, $2\leq p\leq\infty$, 则

$$\log_2 N_2(\Lambda,\varepsilon,m)\leq\left\lceil\frac{c^2d^2}{\varepsilon^2}\right\rceil\log_2(2n+1) \tag{32}$$

在给出定理 1 之前, 先讨论 L_p 范数正则化 SVM 问题, 由非线性最优化理论[42] 知, 式 (9) 可等价为以下约束最小化问题:

$$\min\ \frac{1}{2}\|\boldsymbol{w}\|_p^p+\frac{C}{2}\sum_{i=1}^{m}\left(\max(0,1-y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b))\right)^2 \tag{33}$$

上式可以写为以下约束最小化问题[42]:

$$\min\ \sum_{i=1}^{m}\left(\max(0,1-y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b))\right)^2 \quad \text{s. t.}\ \ \|\boldsymbol{w}\|_p^p\leq B \tag{34}$$

定理 1. 令 $\forall\varepsilon>0$, $\forall\delta>0$, $B>0$, $0<p\leq 2$, 假定支持向量的个数为 N_{SV}, $\xi_M=\max_i\xi_i$, L_p 范数正则化 SVM 算法的抽样复杂性为

$$m=\Omega\left\{\log_2 n\cdot\mathrm{poly}\left(N_{\mathrm{sv}},\xi_M,\frac{1}{\varepsilon},\log_2\frac{1}{\delta}\right)\right\} \tag{35}$$

确保以 $1-\delta$ 概率下式成立:

$$\left|\hat{R}(\boldsymbol{x}_i,y_i;f(x_i))-R(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i))\right|>\varepsilon \tag{36}$$

这里 $\mathrm{poly}(x)$ 表示变量 x 的多项式.

证明. 令

$$\beta(\boldsymbol{w},b)=\sum_{i=1}^{m}\left(\max(0,1-y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b))\right)^2 \tag{37}$$

则

$$\frac{\partial\beta(\boldsymbol{w},b))}{\partial\boldsymbol{w}}=\begin{cases}0, & y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b)\geq 1\\ 2\sum_{i=1}^{m}y_i\boldsymbol{x}_i\left(y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b)-1\right), & y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b)<1-\xi_i\end{cases} \tag{38}$$

根据利普希茨常数的定义,

$$\|\nabla f(\boldsymbol{x}_1)-\nabla f(\boldsymbol{x}_2)\|_2\leq L\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2 \tag{39}$$

假定样例已归一化 $\|\boldsymbol{x}_i\|^2=1$, 满足 $y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b)<1$ 的支持向量的个数为 N_{SV}, 故 $\beta(w,b)$ 的利普希茨常数可由下式求得:

$$\max\frac{\left\|\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}_1}-\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{w}_2}\right\|_2}{\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_2}=\frac{\left\|2\sum_{i=1}^{m}y_i^2\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{w}_1-\boldsymbol{w}_2)\right\|_2}{\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_2}\leq\left\|2\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right\|_2\leq 2N_{\mathrm{SV}} \tag{40}$$

$$M=|\beta(\boldsymbol{w},b)|=\left|\sum_{i=1}^{m}\left(\max(0,1-y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b))\right)^2\right| \tag{41}$$

由于支持向量的样例 $\boldsymbol{x}_i$ 必须满足 $y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b)<1$, 也就是说该样例上的损失不为 0, 又由于间隔带为 $y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b)<\xi_i$, 由定义 $\xi_M=\max_i\xi_i$, 故 $M<N_{\mathrm{SV}}\xi_M^2$.

由引理 3 结合文献 [39] 定理 1 的结果, 知函数类 $F=\left\{f:f=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-b,\ \|\boldsymbol{w}\|_p\leq B\right\}$ 的覆盖数边界为

$$\log_2 N_2\left(F,\varepsilon,m_1\right)\leq\left\lceil\frac{C^2B^2}{\varepsilon^2}\right\rceil\log_2(2n+1) \tag{42}$$

这里 $c=\max_i \|\boldsymbol{x}_i\|_\infty$. 由于 $\|\boldsymbol{x}_i\|_2=1$, 故 $c\le 1$. 结合上述讨论得到:

$$P\left[\exists f\in F:\left|\hat{R}(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i))-R(\boldsymbol{x}_i,y_i;f(\boldsymbol{x}_i))\right|>\varepsilon\right]\le 8(2n+1)2^{128N_{\rm SV}\frac{B^2}{\varepsilon^2}+1}\exp\left(\frac{-m\varepsilon^2}{512N_{\rm SV}^2\xi_M^2}\right) \quad (43)$$

令

$$\delta=8(2n+1)2^{128N_{\rm SV}\frac{B^2}{\varepsilon^2}+1}\exp\left(\frac{-m\varepsilon^2}{512N_{\rm SV}^2\xi_M^2}\right) \quad (44)$$

解出

$$m=\frac{65536N_{\rm SV}^3\xi_M^2B^2In2}{(\varepsilon^2+1)\varepsilon^2}+\frac{512N_{\rm SV}^2\xi_M^2In2}{\varepsilon^2}\log_2\frac{8(2n+1)}{\delta} \quad (45)$$

故 L_p 范数正则化 SVM 算法抽样复杂性为

$$m=\Omega\left\{\log_2{}^n\cdot\mathrm{poly}\left(N_{\rm SV},\xi_M,\frac{1}{\varepsilon},\log_2\frac{1}{\delta}\right)\right\} \quad (46)$$

由定理 1 可知, Lp 范数正则化 SVM 算法抽样复杂性对数依赖于样例的维数, 故能在小样本的情况下, 以高概率实现最优经验风险逼近于期望风险. □

3 实验结果

在第 3.1 节, 我们选择了典型的三个数据集: 前列腺癌 (Prostate-Tumor) 数据集、YaleB0102 人脸数据集和 YaleB2627 人脸数据集对算法的实验结果进行了分析, 为了在足够多的实验数据上验证不同的数据上最优预测性能所在的正则化阶次是不一样的, 在第 3.2 节, 使用了 10 组不同领域的数据并与其他有关算法的预测错误率进行了实验比较, 最后, 在四组基因序列癌症预测数据上与有关的 SVM 分类方法进行了特征选择性能比较.

实验中, 采用种子机制把实验数据划分成训练数据和测试数据, 本实验共有 60 个种子, 能够划分出 60 组不同的训练数据和测试数据, 其中训练数据和测试数据各占 50 %, 总个数为奇数时, 训练数据多分得一个样本. 在上述每个节点上, 每次实验使用一组训练数据, 保证实验数据不会重复; 在整个解路径上, 每个节点的实验数据是相同的, 保证了实验结果的可比性. 在每个节点上针对训练数据的分类和特征选择的结果, 均是通过十倍交叉校验得到.

3.1 三个典型数据集上实验结果

前列腺癌 (Prostate-Tumor) 数据集来自于 http://www.gems-system.org/. 该数据为前列腺癌基因数据, 每个样本包含 10 509 个基因, 共有 102 个样本, 其中癌症样本数为 52, 正常样本数为 50. 本数据属于超高维小样本数据, 对应 10 个不同的 p 值, 算法预测错误率和自动选取特征个数由表 1 给出, 预测错误率和特征个数都为 60 次重复实验的平均值. 从表 1 中可以看出算法的预测错误率在 $p=1.6$, c 为 $(0,1\ 024]$ 中任意值时, 达到最小值 11.57. 在 $p=1.8$ 和 $p=2.0$ 时, 预测错误率显著增大, 主要原因是数据集中, 样例为 10 509 维, 其中很多维包含有噪声, 或者其中一些维与预测类标签无关, $p=1.8$ 和 $p=2.0$ 时, 选中了 10 502.2 和 10 507.3 维, 包含了噪声或无关维变量, 导致模型不准确, 从而降低了预测准确度.

YaleB0102 人脸数据集可在网站 http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html 下载. YaleB 数据为人脸数据, 数据类型为灰度值, 取值范围 0 ~ 255. 该数据包含 38 个人脸数据, 每个人有 59 ~ 64 张人脸照片, 每张照片含 1 024 个像素. 本实验取出第 1 个人和第 2 个人的人脸数据, 共 128 张照片, 其中第 1 个人有 64 张照片, 第 2 个人有 64 张照片. 表 2 为 YaleB0102 数据集预测错误率与特征选取实验结果. 从表 2 中可以看出范数正则化阶次 p 越小, 选取的特征个数越少, 对应于 $p=2.0$ 到 $p=0.2$, 算法选取的特征个数由 1 023.9 下降到 190.3, 算法的预测错误率在 $p=0.2$ 达到最小值 1.88. 在 $p=2.0$ 时, 预测错误率显著增大, 主要原因是数据集中, 样例为 1 024 维, 其中很多维包含有噪声, 或者其中一些维与预测类标签无关, $p=2.0$ 时, 选中了 1 023.9 维, 包含了噪声或无关维变量, 导致模型不准确, 从而降低了预测准确度.

我们还在 YaleB2627 人脸数据做了实验, 限于篇幅没有给出详细结果. 由三个典型数据集实验结果得到当最小分类误差时的特征选择个数及对应的正则化阶次 p 和正则化参数 c 的值, 如表 3 所示.

3.2 分类算法性能比较

首先选择 10 组分别来自人脸图像识别、手写体数字识别、高维文本数据和模型选择挑战数据集, 把提出的算法 (以下简写为 L_p-SVM) 与 L_2-SVM、Zhu 提出的 L_1-SVM、Weston 提出的 L_0-SVM 进行了预测性能的比较研究. Pose05-64X64, PIE-32X32 数据集为人脸图像数据集 (http://www.zjucadcg.cn/dengcai/Data/data.html), MNIST 为手写体数字识别基准数据库数据集 (http://yann.lecun.com/exdb/mnist), Tr11、Tr31、Tr45 和 Hitech 为文本数据集 (http://glaros.dtc.umn.edu/gkhome/cluto/cluto/downloa

d), GINA、GISETTE 和 SYLVA 数据集来自 WCCI2006 模型选择性能预测挑战数据集 (http://www.modelselect.inf.ethz.ch/), 实验的数据划分、数据的特征个数、各算法的参数选择见表 4. L_2-SVM 选择线性核时, L_p-SVM, L_0-SVM, L_1-SVM, L_2-SVM 预测结果见表 5, L_2-SVM 选择径向基核时, 其参数和预测结果见表 6.

表 4～6 中的数据均为 100 次实验的平均结果. 从表 4～6 可以看出, L_2-SVM 使用线性核时, L_p-SVM 优于 L_0-SVM、L_1-SVM 和 L_2-SVM, 且不同的数据上预测误差最小时的 p 值随数据不同而不同. L_2-SVM 使用径向基核时, MNIST, SYLVA 低维小样本数据上, 核映射 L_2-SVM 优于 L_p-SVM; 但是在 Pose05-64×64, PIE-32×32, Tr11, Tr31, Tr45, Hitech, GINA, GISETTE 等高维小样本数据上, L_p-SVM 仍然优于 L_2-SVM.

表 1　Prostate-Tumor 数据集预测错误率与特征选取

Table 1　Prediction error rates and feature slection on Prostate-Tumor dataset

	$p=0.2$	$p=0.4$	$p=0.6$	$p=0.8$	$p=1.0$	$p=1.2$	$p=1.4$	$p=1.6$	$p=1.8$	$p=2.0$
预测错误率 (%)	14.35	13.63	13.33	13.10	12.68	12.32	11.76	11.57	20.39	41.93
特征个数 (个)	119.6	203.7	380.5	902.1	2 522.2	6 544.4	9 744.9	10 428.6	10 502.2	10 507.3

表 2　YaleB0102 数据集预测错误率与特征选取

Table 2　Prediction error rates and feature slection on YaleB0102 dataset

	$p=0.2$	$p=0.4$	$p=0.6$	$p=0.8$	$p=1.0$	$p=1.2$	$p=1.4$	$p=1.6$	$p=1.8$	$p=2.0$
预测错误率 (%)	1.88	1.87	1.92	2.08	2.45	2.92	2.92	2.92	2.76	4.17
特征个数 (个)	190.3	313.3	485.5	680.8	858.9	964.4	1 007.8	1 021	1 023.6	1 023.9

表 3　三个典型数据集上最小分类误差时的参数值及特征选择个数

Table 3　Parameter values and numbers of feature selection on three representative datasets

数据名	正则化阶次 p	正则化参数 c	最小分类误差 (%)	特征选择个数
Prostate-Tumor	1.6	(0,1 024]	11.57	10 428.6
YaleB0102	0.6	$c \geq 2$	1.88	190.3
YaleB2627	0.6	$0.125 < c \leq 1\,024$	4.72	510.6

表 4　10 个数据集的特性及 L_p-SVM 与 L_0-SVM, L_1-SVM, L_2-SVM 算法参数

Table 4　Properties of ten datasets and parameter setups of L_p-SVM, L_0-SVM, L_1-SVM, and L_2-SVM algorithm

数据名称	训练 + 测试	特征个数	L_p 参数 $(c+p)$	L_0 参数 c	L_1 参数 c	L_2 参数 c
Pose05-64×64	49 + 49	4 096	2 + 1.8	10	10	10
PIE-32×32	170 + 170	1 024	0.5 + 0.4	10	10	10
MNIST	500 + 300	784	0.0625 + 1.6	10	10	10
Tr11	72 + 71	6 424	0.25 + 1.8	10	10	10
Tr31	189 + 189	10 127	0.5 + 1.2	10	10	10
Tr45	144 + 144	8 261	16 + 0.4	10	10	10
Hitech	483 + 483	10 080	0.0625 + 1.4	10	10	10
GINA	100 + 100	970	0.125 + 0.2	10	10	10
GISETTE	200 + 200	5 000	8 + 1.4	10	10	10
SYLVA	654 + 654	216	0.25 + 0.2	10	10	10

表 7 给出了 10 个数据集上 L_p-SVM, L_0-SVM, L_1-SVM 和 L_2-SVM 分类算法使用上述模型选择过程得到的最优参数值时训练时间的比较结果, 表中是 100 次运行时间的平均结果, 时间单位为秒. 可以看出, 我们提出的 L_p-SVM 在 PIE-32×32, MNIST, Hitech, GINA 和 SYLVA 数据集上运行时间比其他两个具有特征选择功能的分类算法 L_0-SVM, L_1-SVM 用时少, 在其余数据集上训练时间与 L_0-SVM, L_1-SVM 在同一数量级上, 三个具有特征选择功能的分类算法的训练时间都比 L_2-SVM 分类算法用时时间长.

为了验证特征选择的正确率, 我们把 L_p-SVM 与 L_2-SVM, L_1-SVM, L_0-SVM 进行了特征选择性能比较, L_p-SVM, L_1-SVM 和 L_0-SVM 三个算法均没有使用核映射. 我们选择了 4 组癌症预测数据: 乳腺癌数据 (http://www.ihes.fr/zinovyev/princmanif 2006/) 由 42 个训练样本 15 个测试样本组成, 共 57 个样本, 每个样本的维数为 2 215; 黑素瘤数据 (http://www.cancerinstitute.org.au/cancerinst /nswog/groups/melanoma1.html) 由 58 个训练样本 20 个测试样本组成, 每个样本的维数为 3 750; 淋巴癌数据 (http://llmpp.nih.gov/ lymphoma/data/rawdata/) 由 72 个训练样本 24 个测试样本组成, 共 96 个样本, 每个样本的维数为 4 026; 结肠癌数据 (http://perso.telecom-paristech.fr/gfort/GLM/Programs.html) 由 46 个训练样本、16 个测试样本组成, 共 62 个样本, 每个样本的维数为 2 000. 对于癌症数据中所包含的基因阵列, 生物学家已经标注出哪些基因与癌症有关, 如果预测算法选择了与该癌症类型有关的基因组, 则认为特征选择正确. 实验的参数选择使用十倍交叉校验过程得到, 最终得到的实验结果如表 8 所示. 表 8 中所有预测值均是 100 次实验的平均值. 可以看出, L_p-SVM 在 4 组癌症数据集上的预测错误率是五种方法最低的, 实现了准确的基因选择. 而 L_2-SVM 由于算法训练得到的所有分量均不为零, 亦即选择了样本的所有坐标参与模型构造, 选择了所有的特征, 不能分辨哪些基因与相应的癌症有关.

表 5 10 个数据集上 L_p-SVM 与 L_0-SVM, L_1-SVM, 线性核 L_2-SVM 预测结果

Table 5 Prediction results of L_p-SVM, L_0-SVM, L_1-SVM and linear kernel L_2-SVM on ten datasets

数据名称	原始标签	L_p 预测错误率 (%)	L_0 预测错误率 (%)	L_1 预测错误率 (%)	L_2 预测错误率 (%)
Pose05-64×64	25, 36	0	0.31	0.71	17.45
PIE-32×32	5, 8	0.06	2.09	0.35	0.88
MNIST	2, 7	0.9	1.45	0.97	0.93
Tr11	2, 6	0.14	1.97	0.56	0.21
Tr31	3, 5	1.3	2.22	1.43	1.43
Tr45	1, 5	1.18	2.19	1.42	1.94
Hitech	1, 5	3.93	4.32	1.44	4.29
GINA	−1, 1	13.96	22.1	18.61	18.97
GISETTE	−1, 1	8.42	10.53	8.95	9.43
SYLVA	−1, 1	1.54	2.58	2.13	2.42

表 6 10 个数据集上核映射 L_2-SVM 的预测结果及参数

Table 6 Prediction results and parameter setup of L_2-SVM using nonlinear kernel map on ten datasets

数据名称	预测错误率 (%)	参数 c	核参数	核类型
Pose05-64×64	0.43	10	10	径向基核
PIE-32×32	1.39	10	10	径向基核
MNIST	0.62	10	10	径向基核
Tr11	1.49	10	85	径向基核
Tr31	3.24	150	62	径向基核
Tr45	1.61	10	77	径向基核
Hitech	4.38	10	85	径向基核
GINA	14.94	10	10	径向基核
GISETTE	10.4	10	20	径向基核
SYLVA	0	0.5	50	径向基核

表 7　10 个数据集上 L_p-SVM 与 L_0-SVM, L_1-SVM, 线性核 L_2-SVM 运行时间比较

Table 7　Comparison of running time of L_p-SVM, L_0-SVM, L_1-SVM, and linear kernel L_2-SVM on ten datasets

数据名称	原始标签	L_p-SVM 耗时 (s)	L_0-SVM 耗时 (s)	L_1-SVM 耗时 (s)	L_2-SVM 耗时 (s)
Pose05-64×64	25, 36	1.26	0.39	1.58	0.14
PIE-32×32	5, 8	0.57	6.39	2.08	0.13
MNIST	2, 7	1.90	65.80	60.66	0.59
Tr11	2, 6	2.21	0.86	0.47	0.26
Tr31	3, 5	8.49	9.07	4.06	2.40
Tr45	1, 5	5.15	4.17	1.67	1.19
Hitech	1, 5	22.72	88.90	58.18	10.15
GINA	-1, 1	0.32	1.71	0.44	0.06
GISETTE	-1, 1	4.33	10.76	3.34	1.19
SYLVA	-1, 1	2.34	99.70	106.36	0.38

表 8　四个癌症数据集四种分类算法预测错误率和特征选择错误率比较 (%)

Table 8　Comparison of the prediction error rates and feature selection error rates of four classifiers on four cancer datasets (%)

错误率	数据名称	L_p-SVM	L_2-SVM	L_1-SVM	L_0-SVM
预测错误率	乳腺癌数据	20.1	23.6	26.41	28.41
	黑素瘤数据	13.68	14.71	14.11	18.13
	淋巴癌数据	6.67	7.39	6.10	10.10
	结肠癌数据	10	12.2	16.39	11.48
特征选择错误率	乳腺癌数据	3.47	–	4.21	9.28
	黑素瘤数据	0.25	–	1.5	6.34
	淋巴癌数据	0.11	–	1.62	5.71
	结肠癌数据	2.34	–	3.10	8.21

4　结论与展望

SVM 为当前主流机器学习算法. 以前的 SVM 分类器事先假定正则化的阶次为 1 范数或者 2 范数. 当前非常重视同时实现精确分类和特征选择的分类器设计模式. 我们通过大量的实验研究发现, 不同的数据集适宜于选择不同的正则化阶次, 有理数范数正则化能够实现特征选择.

本文提出了 p 范数正则化 SVM 分类器算法设计模式, 算法能够选择正则化范数的阶次 p, $0 < p \leq 2$, 并提出了求解 p 范数正则化 SVM 问题的迭代再权算法. 根据预测误差 $err(c,p)$ 性能指标, 我们设计了解路径寻踪网格算法, 在算法的两个可调参数 (正则化参数 c 和正则化阶次 p) 的取值范围内, 依次求解 p 范数正则化 SVM 目标函数. 在实验部分, 使用三组有代表性的数据集 Prostate-Tumor 数据、YaleB0102 人脸数据和 YaleB2627 人脸数据集, 对提出的算法进行了实验研究, 并使用不同领域 10 组数据进行了预测性能的比较, 实验结果表明, 在最小分类误差处, 正则化阶次 p 不是 0, 1, 或者 2, 而是某个非整数值, p 范数正则化 SVM 分类器优于其他线性正则化分类器. 最后, 选择四组基因序列癌症预测数据, 把提出的分类算法与 L_2-SVM, L_1-SVM, L_0-SVM 进行了特征选择实验比较, 证明提出的算法的特征选择错误率均小于其他算法, 实验结果表明了提出算法的优越性.

结论表明: 在设计 SVM 分类算法时, 强制性的采用整数范数, 分类效果并不一定好. 采用有理数范数, 分类效果往往比整数范数分类效果好, 该有理数的具体取值应该由算法根据不同数据自行确定.

本文的工作无法实现数据无穷维核映射后的分类器设计, 提出的迭代再权算法每一次迭代需要 $\boldsymbol{w}$ 的显式表达, 特征选择也需要 $\boldsymbol{w}$ 的显式表达, 如何实现无穷维核空间的 p 范数正则化 SVM, 是一个值得考虑的问题, 另外, p 范数正则化 SVM 可以拓展

到硬间隔、一次软间隔 SVM 分类器设计中, 这是我们下一步要做的工作.

References

1 Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh, USA: ACM, 1992. 144−152

2 Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, **20**(3): 273−297

3 Weston J, Elisseeff A, Scholkopf B, Tipping M. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 2003, **3**: 1439−1461

4 Liu Qiao, Qin Zhi-Guang, Chen Wei, Zhang Feng-Li. Zero-norm penalized feature selection support vector machine. *Acta Automatica Sinica*, 2011, **37**(2): 252−256
(刘峤, 秦志光, 陈伟, 张凤荔. 基于零范数特征选择的支持向量机模型. 自动化学报, 2011, **37**(2): 252−256)

5 Liu Z, Jiang F, Tian G, Wang S, Sato F, Meltzer S J, et al. Sparse logistic regression with L_p penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology*, 2007, **6**(1): 1−20

6 Shi J N, Yin W T, Osher S, Sajda P. A fast hybrid algorithm for large-scale L_1-regularized logistic regression. *Journal of Machine Learning Research*, 2010, **11**: 713−741

7 Liu Y F, Wu Y C. Variable selection via a combination of the L_0 and L_1 penalties. *Journal of Computational and Graphical Statistics*, 2007, **16**(4): 782−798

8 Liu Y F, Zhang H H, Park C, Ahn J. Support vector machines with adaptive L_q penalties. *Computational Statistics and Data Analysis*, 2007, **51**(12): 6380−6394

9 Mangasarian O L, Gang K. Feature selection for nonlinear kernel support vector machines. In: Proceedings of the 7th IEEE International Conference on Data Mining Workshops. Omaha, USA: IEEE, 2007. 231−236

10 Foucart S, Lai M J. Sparsest solutions of under determined linear system via L_q-minimization for $0 < q < 1$. *Applied and Computational Harmonic Analysis*, 2009, **26**(3): 395−407

11 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 2001, **96**(456): 1348−1360

12 Holland P W, Welsch R E. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, 1977, **6**(9): 813−827

13 Gorodnitsky I F, Rao B D. Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 1997, **45**(3): 600−616

14 Daubechies I, DeVor R, Fornasier M, Gunturk C S. Iteratively re-weighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 2008, **63**(1): 1−38

15 Chartrand R, Yin W. Iteratively reweighted algorithms for compressive sensing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA: IEEE, 2008. 3869−3872

16 Candes E J, Wakin M B, Boyd S P. Enhancing sparsity by reweighted L_1 minimization. *Journal of Fourier Analysis and Applications*, 2008, **14**(5−6): 877−905

17 Candes E J, Wakin M B, Boyd S P. Enhancing sparsity by reweighted L_1 minimization. *Journal of Fourier Analysis and Applications*, 2008, **14**(5): 877−905

18 Wang L, Zhu J, Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 2008, **24**(3): 412−419

19 Wang L, Zhu J, Zou H. The doubly regularized support vector machine. *Statistica Sinica*, 2006, **16**(2): 589−615

20 Wechsler H. *Reliable Face Recognition Methods: System Design, Implementation and Evaluation*. New York: Springer, 2010

21 Liu D H, Lam K M, Shen L S. Illumination invariant face recognition. *Pattern Recognition*, 2005, **38**(10): 1705−1716

22 Mian A. Online learning from local features for video-based face recognition. *Pattern Recognition*, 2011, **44**(5): 1068−1075

23 Zhang Xue-Gong. Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica*, 2000, **26**(1): 32−42
(张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, **26**(1): 32−42)

24 Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines. In: Proceedings of the IEEE Workshop Neural Networks for Signal Processing. Amelia Island, USA: IEEE, 1997. 276−285

25 Platt J C. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods: Support Vector Learning*. Cambridge: MIT Press, 1999. 185−208

26 Joachims T. Making large-scale support vector machine learning practical. *Advances in Kernel Methods: Support Vector Learning*. Cambridge: MIT Press, 1999. 169−184

27 Shalev-Shwartz S, Singer Y, Srebro N. Pegasos: primal estimated sub-gradient solver for SVM. In: Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA: ACM, 2007. 807−814

28 Hsieh C J, Chang K W, Lin C J, Keerthi S S, Sundararajan S. A dual coordinate descent method for large-scale linear SVM. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 408−415

29 Teo C H, Vishwanthan S V N, Smola A J, Le Q V. Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 2010, **11**: 311−365

30 Mangasarian O L. Exact 1-norm support vector machines via unconstrained convex differentiable minimization. *Journal of Machine Learning Research*, 2006, **7**: 1517−1530

31 Bradley P S, Mangasarian O L. Feature selection via concave minimization and support vector machines. In: Proceedings of the 15th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 1998. 82−90

32 Zhang H H, Liu Y, Wu Y, Zhu J. Variable selection for multicategory SVM via sup-norm regularization. *Electronic Journal of Statistics*, 2008, **2**: 149−167

33 Wang L, Shen X. On L_1-norm multiclass support vector machines: methodology and theory. *Journal of the American Statistical Association*, 2007, **102**(478): 583−594

34 Amaldi E, Kann V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 1998, **209**(1−2): 237−260

35 Scholkopf B, Smola A J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press. 2001

36 Ng A Y. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In: Proceedings of the 21st International Conference on Machine Learning. Banff, Canada: ACM, 2004. 1−8

37 Lorentz G G. Metric entropy and approximation. *Bulletin of the American Mathematical Society*, 1966, **72**(6): 903−937

38 Kolmogorov A N, Tikhomirov V M. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations*, 1961, **17**(2): 277−364

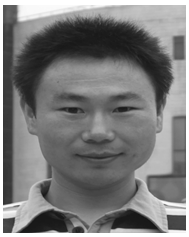
39 Kaban A, Durrant R J. Learning with $L_q < 1$ vs. L_1-norm regularisation with exponentially many irrelevant features. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Antwerp, Belgium: Springer, 2008. 580−596

40 Pollard D. *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984

41 Zhang T. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2002, **2**: 527−550

42 Luenberger D G, Ye Y Y. *Linear and Nonlinear Programming (Third Edition)*. Boston: Springer, 2007

刘建伟 中国石油大学 (北京) 地球物理与信息技术学院自动化系副研究员. 主要研究方向为智能信息处理, 复杂系统分析, 预测与控制, 算法分析与设计. 本文通信作者.
E-mail: liujw@cup.edu.cn
(**LIU Jian-Wei** Associate professor in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing Campus. His research interest covers intelligent information processing, machine learning, analysis, prediction, controlling of complicated nonlinear system, and analysis of the algorithm and the designing. Corresponding author of this paper.)



李双成 中国石油大学 (北京) 地球物理与信息技术学院自动化系硕士研究生. 主要研究方向为机器学习与数字图像处理. E-mail: shchli2003@163.com
(**LI Shuang-Cheng** Master student in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing Campus. His research interest covers machine learning and digital image processing.)

罗雄麟 中国石油大学 (北京) 地球物理与信息技术学院自动化系教授. 主要研究方向为智能控制, 复杂系统分析, 预测与控制. E-mail: luoxl@cup.edu.cn
(**LUO Xiong-Lin** Professor in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing Campus. His research interest covers intelligent control, and analysis, prediction, controling of complicated nonlinear system.)