第 46 卷 第 12 期 计 算 机 工 程 2020 年 12 月
Vol. 46 No. 12 Computer Engineering December 2020

·人工智能与模式识别· 文章编号：1000-3428(2020)12-0088-08 文献标志码：A 中图分类号：TP181

基于动量加速零阶减小方差的鲁棒支持向量机

鲁淑霞，蔡莲香，张罗幻

（河北大学 数学与信息科学学院 河北省机器学习与计算智能重点实验室，河北 保定 071002）

摘 要： 在实际分类问题中，由于人为或其他因素的影响，数据中往往存在一定的噪声，而传统支持向量机（SVM）使用的铰链损失函数对噪声数据敏感，且分类性能较差。为消除噪声数据的影响，提出一种新的鲁棒 SVM 算法。通过引入新形式的损失函数，并基于间隔分布的思想，建立鲁棒 SVM 优化模型提高 SVM 的抗噪性，运用零阶减小方差算法并结合动量加速技术，给出一种新的优化模型求解方法。实验结果表明，该方法通过引入梯度修正项降低了方差对算法的影响，同时结合动量加速技术，明显提高了算法的收敛速度。

关键词： 噪声；零阶梯度；方差；动量加速；鲁棒支持向量机

开放科学（资源服务）标志码（OSID）：

中文引用格式： 鲁淑霞，蔡莲香，张罗幻. 基于动量加速零阶减小方差的鲁棒支持向量机[J]. 计算机工程，2020，46(12)：88-95，104.

英文引用格式： LU Shuxia，CAI Lianxiang，ZHANG Luohuan. Robust support vector machine based on momentum acceleration zeroth-order variance reduction[J]. Computer Engineering，2020，46(12)：88-95，104.

Robust Support Vector Machine Based on Momentum Acceleration Zero-Order Variance Reduction

LU Shuxia，CAI Lianxiang，ZHANG Luohuan

（Hebei Province Key Laboratory of Machine Learning and Computational Intelligence，College of Mathematics and Information Science，Hebei University，Baoding，Hebei 071002，China）

【Abstract】 In the actual classification problem，there is often a certain amount of noise in the data caused by the influence of artificial or other factors，so it is very important to improve the anti-noise ability of the classifier. However，the hinge loss function used by the traditional Support Vector Machine（SVM） is sensitive to noisy data and has poor classification performance. In order to eliminate the influence of noisy data，this paper proposes a robust SVM based on momentum acceleration zero-order variance reduction. By introducing a new form of loss function and adopting the idea of margin distribution，a robust SVM optimization model is established to improve the anti-noise ability of SVM. By using the zero-order variance reduction algorithm and momentum acceleration technique，a new optimization model solution method is proposed. Experimental results show that this method reduces the influence of variance effectively by introducing the gradient correction item，and increases the convergence speed of the algorithm significantly by using the momentum acceleration technology.

【Key words】 noise；zero-order gradient；variance；momentum acceleration；robust Support Vector Machine（SVM）

DOI： 10.19678/j.issn.1000-3428.0056286

0 概述

间隔理论[1]由 VAPNIK 等人于 1995 年提出，其基于最小间隔最大化的原理，能够从理论上有效地解释其他许多学习方法的泛化性。在间隔理论的基础上，文献[2]提出了一种类似于 boosting 的算法 Arc-gv，该算法同样以最小间隔最大化的方式求解优化问题，但泛化性能较差。研究人员发现这种算法虽然充分利用了最小间隔的重要性，但是数据的间隔分布并不好。他们认为相比于最小间隔，数据的间隔分布可能对泛化性的影响更大，文献[3]对其进行了理论证明。并且，文献[4]还将该思想应用到

基金项目： 国家自然科学基金（61672205）。

作者简介： 鲁淑霞（1966—），女，教授、博士，主研方向为机器学习；蔡莲香、张罗幻，硕士研究生。

收稿日期： 2019-10-14 **修回日期：** 2019-11-19 **E-mail：** cmclusx@126.com

传统支持向量机(Support Vector Machine,SVM)分类中,获得了更好的分类精度和泛化性能,充分说明相比传统的最小间隔最大化的优化方法,对间隔分布进行优化更加重要。

在现实的分类问题中,由于人为或其他因素的影响,数据往往会存在一定的噪声。如何对带有噪声的数据进行有效分类,是一个值得研究的问题。然而,传统的 SVM 对噪声数据不具有很好的鲁棒性,这是因为传统 SVM 使用无界的铰链损失函数,对于噪声数据会产生较大的损失值,使 SVM 的分类超平面严重偏离最优超平面,影响最终的分类效果。于是,许多研究从改进损失函数角度出发,提高 SVM 对噪声数据的鲁棒性。文献[5]在铰链损失的基础上提出了一种截断的铰链损失,通过引入一个小于 0 的截断参数 s,使铰链损失有一个确定的界限,解决了噪声数据带来较大损失的问题。通过最大化 2 个类之间的最小分位数距离,文献[6]提出了弹球损失,弹球损失是最大化 2 个类之间的分位数间距,而不是最小间距,由此提高了对属性噪声的识别性,改善了分类性能。文献[6]在此基础上提出了一种截断的弹球损失(truncated pinball loss)。相比于最初的 pinball 损失,文献[7]提出的损失函数增加了两段水平部分,使得损失函数的值有一个固定的上界,降低了噪声数据对算法性能的影响。

求解建立的 SVM 模型也是一个值得研究的问题。随机梯度下降(Stochastic Gradient Descent,SGD)算法是一阶优化方法,广泛应用于各种优化问题中,并衍生出许多优秀的算法,如文献[8]提出的 Pegasos 算法,该算法在每次迭代中随机选择一个样本计算梯度,并以此代替全梯度。由于通常假设样本是独立同分布的,从而随机抽取单个样本的目标函数的梯度是整个目标函数梯度的无偏估计,进而可用每次迭代仅处理单个或部分样本的随机优化方法来代替批处理方法,但是该方法存在方差,随着迭代次数的增加,方差也逐渐累加,收敛速率不可避免地受到影响。为降低方差的影响,文献[9]提出了减小方差的随机梯度(Stochastic Variance Reduction Gradient,SVRG)下降算法。该算法分为内外两层循环,仅在外层循环计算全梯度,降低了计算量。在内层循环中引入梯度修正项,降低了方差对算法的影响,提高了算法的性能。文献[10]在 SVRG 算法的基础上依据 Nesterov 的动量加速技巧,提出了快速减小方差的随机梯度(Fast Stochastic Variance Reduced Gradient,FSVRG)下降[11]算法。FSVRG 是 SVRG 的一个加速变种,在每次内层迭代中引入了动量加速技巧,不仅计算在当前迭代中的梯度值,同时考虑了上一轮的梯度变化,与 SVRG 相比提高了算法的收敛速度。此外,在文献[12]提出的结构凸优化问题和文献[13]提出的经典的 Katyusha 算法以及文献[14]提出的加速随机镜像下降算法中均引入了动量加速技巧,且都获得了较好的性能。

上述所提方法虽然可以有效地降低方差对算法的影响,但是在每次迭代中均需要计算梯度,求解梯度困难或者不能求解梯度的模型,会增加额外的开销。因此,文献[15]把零阶优化方法和减小方差策略相结合,提出一种零阶减小方差的随机梯度(Zeroth Order-Stochastic Variance Reduced Gradient,ZO-SVRG)下降方法。ZO-SVRG 不需要计算梯度的准确数值,而是用函数值逼近梯度,有效地解决了复杂模型不能计算梯度的问题,具有较高的实用性。文献[16]基于零阶优化的思想,结合交替方向乘子法,提出一种在线的零阶交替方向乘子算法(ZOO-ADMM)。该算法既避免了梯度的计算,又利用了交替乘子法能够处理复杂结构的优势,经过理论分析和实验验证说明了所提方法的有效性。

为解决传统 SVM 对噪声敏感的问题,本文通过引入间隔分布和新形式的损失函数,提出一种基于动量加速零阶减小方差的鲁棒支持向量机(MA-ZOVR)。

1　相关工作

对于一般的优化问题,定义如下的样例空间:$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ 为给定的 n 个数据,$x_i\in\mathbb{R}^d$,d 表示样例的维数,$y=\{-1,+1\}$ 表示样例标签,则传统的 SVM 的优化问题可以表示为下述形式:

$$\min F(\boldsymbol{\omega})=\frac{\lambda}{2}\|\boldsymbol{\omega}\|^2+\frac{1}{n}\sum_{i=1}^{n}\max(0,1-y_i\boldsymbol{\omega}^{\mathrm{T}}x_i) \tag{1}$$

其中,$\frac{\lambda}{2}\|\boldsymbol{\omega}\|^2$ 是目标函数的正则化项。

1.1　零阶减小方差算法

在零阶优化方法中,对于优化问题的梯度计算不再使用传统的方法,而是直接使用函数值近似代替梯度,避免了重复的梯度计算,大幅降低了工作量。对于计算梯度困难的复杂模型,零阶优化可以有效地提高算法性能。因此,零阶优化广泛应用于机器学习领域,并由此衍生出一系列算法。具有代表性的零阶优化算法是文献[12]提出的 ZO-SVRG 算法。ZO-SVRG 算法分为内外双重循环,在零阶优化的基础上结合了 SVRG 减小方差的策略。在每轮外层循环中,定义一个向量 $\tilde{\boldsymbol{\omega}}$,该向量是上一轮进行 T 次内层迭代后的最后一个值,同时仅在每轮外层循环中计算所有样本在 $\tilde{\boldsymbol{\omega}}$ 处的全梯度 $\hat{\nabla}F(\tilde{\boldsymbol{\omega}})$。在 ZO-SVRG 算法的 T 次内层迭代中每次随机抽取一个样本 i,计算单个样本梯度 $\hat{\nabla}F_i(\boldsymbol{\omega}_t)$ 及单个样本在 $\tilde{\boldsymbol{\omega}}$ 处的梯度 $\hat{\nabla}F_i(\tilde{\boldsymbol{\omega}})$,并进行如下的梯度修正:

$$G(\boldsymbol{\omega}_t)=\hat{\nabla}F_i(\boldsymbol{\omega}_t)-\hat{\nabla}F_i(\tilde{\boldsymbol{\omega}})+\hat{\nabla}F(\tilde{\boldsymbol{\omega}}) \quad (2)$$

其中,$G(\boldsymbol{\omega}_t)$称为梯度修正项。

零阶梯度估计用函数值近似代替,坐标梯度估计如下[15]:

$$\hat{\nabla}F_i(\boldsymbol{\omega})=\sum_{l=1}^{d}\left(\frac{1}{2\mu_l}\right)[F_i(\boldsymbol{\omega}+\mu_l\boldsymbol{e}_l)-F_i(\boldsymbol{\omega}-\mu_l\boldsymbol{e}_l)]\boldsymbol{e}_l \quad (3)$$

其中,$\boldsymbol{e}_l$ 表示一个标准基向量,在第 l 个坐标处为 1,其他坐标处为 0,$\mu_l>0$ 表示光滑参数。

ZO-SVRG 算法如下:

算法 1 ZO-SVRG 算法

输入 外层迭代轮数 S,内层迭代次数 T,学习率 η,光滑参数 μ

输出 $\tilde{\boldsymbol{\omega}}_s$

初始化 $\tilde{\boldsymbol{\omega}}_0$

for s = 1,2,…,S

  $\tilde{\omega}=\tilde{\omega}_{s-1}$

  计算全梯度 $\hat{\nabla}F(\tilde{\omega})$

  $\omega_0=\tilde{\omega}$

  for t = 0,1,…,T − 1

    随机抽取一个样本 i,进行梯度更新

    $\omega_{t+1}=\omega_t-\eta(\hat{\nabla}F_i(\omega_t)-\hat{\nabla}F_i(\tilde{\omega})+\hat{\nabla}F(\tilde{\omega}))$

  end

  $\tilde{\omega}_s=\omega_T$

end

return $\tilde{\omega}_s$

1.2 动量加速技巧

文献[17]指出 Nesterov[10] 的动量加速技巧可以加速随机减小方差算法的收敛,能够使强凸问题和一般凸问题的收敛速度达到较高的水平。随机减小方差算法均分为内外双重循环,引入动量加速技巧后,每次内层迭代的梯度由式(4)、式(5)2 个更新规则构成:

$$v_{t+1}=v_t-\eta_s(\hat{\nabla}F_i(\boldsymbol{\omega}_t)-\hat{\nabla}F_i(\tilde{\boldsymbol{\omega}})+\hat{\nabla}F(\tilde{\boldsymbol{\omega}})) \quad (4)$$

$$\boldsymbol{\omega}_{t+1}=\tilde{\boldsymbol{\omega}}+\rho_s(v_{t+1}-\tilde{\boldsymbol{\omega}}) \quad (5)$$

根据式(4)计算当前迭代中的梯度变化情况,其中,v_{t+1} 为辅助变量,η_s 为更新步长。

式(5)表示,结合上一轮梯度的结果,得出本次内层迭代最终的梯度更新规则。其中,ρ_s 表示动量权重系数。

引入动量加速技巧后的梯度更新规则并不仅仅依赖于当前迭代的梯度变化情况,并且考虑了上一轮的最终结果。所以,计算当前迭代中的梯度时,都会有一个之前梯度的作用。如果这次的梯度和上一轮的梯度方向相同,则会因为之前的速度继续加速;如果这次的梯度和上一轮的梯度方向相反,则不增加或减少过多。因此,引入动量加速技巧后,会使梯度在每次的下降过程中减少摆动的幅度,加速算法的收敛。

2 动量加速零阶减小方差的鲁棒 SVM

基于文献[4]的理论证明,在本文中用间隔分布均值代替传统的最小间隔,充分考虑每个样本的分布情况。间隔分布均值为:

$$\frac{1}{n}\sum_{i=1}^{n}y_i\boldsymbol{\omega}^{\mathrm{T}}x_i \quad (6)$$

此外,传统 SVM 的损失函数为铰链损失。但由于铰链函数无界性的特点,对于噪声数据会带来较大的损失,使分类性能下降。因此,为了降低噪声数据的影响,结合指数函数的结构特点,引入一种新形式的损失函数:

$$\max\left(0,\frac{\mathrm{e}^{2(1-y_i\boldsymbol{\omega}^{\mathrm{T}}x_i)}-1}{\mathrm{e}^{2(1-y_i\boldsymbol{\omega}^{\mathrm{T}}x_i)}+1}\right) \quad (7)$$

损失函数曲线如图 1 所示。

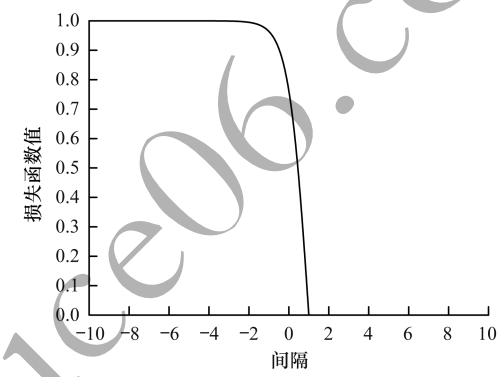

图 1 损失函数曲线示意图

Fig. 1 Schematic diagram of loss function curve

在图 1 中,横坐标表示间隔,即 $y_i\boldsymbol{\omega}^{\mathrm{T}}x_i$,一般又将其称为函数间隔。在 SVM 的分类任务中,可以通过函数间隔的正负来判定或表示样本分类的正确性;纵坐标表示对应的损失函数值。下文对损失函数进行具体的分析:

1)当 $1-y_i\boldsymbol{\omega}^{\mathrm{T}}x_i=0$ 时,$y_i\boldsymbol{\omega}^{\mathrm{T}}x_i=1>0$,样本正确分类,对应的损失值为 0。

2)当 $1-y_i\boldsymbol{\omega}^{\mathrm{T}}x_i<0$ 时,$y_i\boldsymbol{\omega}^{\mathrm{T}}x_i>1>0$,样本正确分类,对应的损失值为 0。

3)当 $1-y_i\boldsymbol{\omega}^{\mathrm{T}}x_i>0$ 时,$y_i\boldsymbol{\omega}^{\mathrm{T}}x_i<1$,这时存在 2 种情况:

(1)当 $0<y_i\boldsymbol{\omega}^{\mathrm{T}}x_i<1$ 时,样本位于分类间隔内。

(2)当 $y_i\boldsymbol{\omega}^{\mathrm{T}}x_i<0$ 时,样本错误分类。

在 3)中的 2 种情况表示的是噪声数据的分类状态。可以明显地看出噪声数据的损失值限制在了[0,1]之间,避免了出现噪声数据带来过大损失的情况,从而降低了噪声数据对分类性能的影响。

2.1　线性 MA-ZOVR 算法

通过引入间隔分布均值以及新的损失函数,在线性输入空间建立新的优化模型如下:

$$\min F(\boldsymbol{\omega}) = -\frac{\lambda}{n}\sum_{i=1}^{n} y_i \boldsymbol{\omega}^{\mathrm{T}} x_i + \frac{1}{n}\sum_{i=1}^{n}\max\left(0, \frac{\mathrm{e}^{2(1-y_i\boldsymbol{\omega}^{\mathrm{T}}x_i)}-1}{\mathrm{e}^{2(1-y_i\boldsymbol{\omega}^{\mathrm{T}}x_i)}+1}\right) \quad (8)$$

对于优化模型的求解,基于零阶优化减小方差的思想,配合动量加速技巧,提出一种基于动量加速零阶减小方差(MA-ZOVR)算法。

在线性 MA-ZOVR 算法中,首先对梯度进行估计,改变传统的梯度计算方式,通过式(3)坐标梯度估计法,计算出函数值并近似代替梯度。为降低方差对算法性能的影响,引入了式(2)的梯度修正项。最后结合动量加速技巧,在内层迭代中使用式(4)和式(5)进行梯度的更新。

线性 MA-ZOVR 算法如算法 2 所示。

算法 2　线性 MA-ZOVR 算法

输入　外层迭代轮数 S,内层迭代次数 T,光滑参数 μ,正则化参数 λ

输出　$\hat{\boldsymbol{\omega}}_S$

初始化　$v_0 = \tilde{\boldsymbol{\omega}}_0$,$\{\rho_s\}$,$\beta > 0$,$\eta_0$

1. for s = 1,2,⋯,S
2. $\tilde{\omega} = \tilde{\omega}_{s-1}$
3. 计算全梯度 $\hat{\nabla} F(\tilde{\omega})$
4. $\eta_s = \dfrac{\eta_0}{\max\left(\beta, \dfrac{2}{s+1}\right)}$
5. $\omega_0 = \rho_s v_0 + (1-\rho_s)\tilde{\omega}$
6. for t = 0,1,⋯,T − 1
7. 随机抽取一个样本 i 进行梯度更新
8. $v_{t+1} = v_t - \eta_s(\hat{\nabla} F_i(\omega_t) - \hat{\nabla} F_i(\tilde{\omega}) + \hat{\nabla} F(\tilde{\omega}))$
9. $\omega_{t+1} = \tilde{\omega} + \rho_s(v_{t+1} - \tilde{\omega})$
10. end for
11. $\tilde{\omega}_s = \dfrac{1}{T}\sum_{t=1}^{T}\omega_t$
12. $v_0 = v_T$
13. end for
14. $\hat{\omega}_s = \tilde{\omega}_s$, if $F(\tilde{\omega}_S) \leqslant F\left(\dfrac{1}{S}\sum_{s=1}^{S}\tilde{\omega}_s\right)$,

$\hat{\omega}_S = \dfrac{1}{S}\sum_{s=1}^{S}\tilde{\omega}_s$, otherwise

算法中的第 8 步、第 9 步为梯度更新规则,也是关键的步骤。其中 $\rho_s = \max\left(\beta, \dfrac{2}{s+1}\right)$ 表示动量权重[7]。

2.2　非线性 MA-ZOVR 算法

在非线性输入空间,定义如下的优化问题:

$$\min F(\boldsymbol{\omega}) = -\frac{\lambda}{n}\sum_{i=1}^{n} y_i \boldsymbol{\omega}^{\mathrm{T}} \varphi(x_i) + \frac{1}{n}\sum_{i=1}^{n}\max\left(0, \frac{\mathrm{e}^{2(1-y_i\boldsymbol{\omega}^{\mathrm{T}}\varphi(x_i))}-1}{\mathrm{e}^{2(1-y_i\boldsymbol{\omega}^{\mathrm{T}}\varphi(x_i))}+1}\right) \quad (9)$$

一般地,在非线性特征空间,优化问题中 $\varphi(x_i)$ 的维数很高,求解非常复杂。本文通过表示定理[8,20]对式(9)进行变形:

$$\boldsymbol{\omega} = \sum_{i=1}^{n}\alpha_i\varphi(x_i) = \boldsymbol{X}\boldsymbol{\alpha} \quad (10)$$

其中,$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_n]^{\mathrm{T}}$,$\boldsymbol{X} = [\varphi(x_1), \varphi(x_2), \cdots, \varphi(x_n)]$。根据式(10)可得:

$$y_i\boldsymbol{\omega}^{\mathrm{T}}\varphi(x_i) = y_i(\boldsymbol{X}\boldsymbol{\alpha})^{\mathrm{T}}\varphi(x_i) = y_i\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\varphi(x_i) = y_i\boldsymbol{\alpha}^{\mathrm{T}}G_i$$

其中,$\boldsymbol{G} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 表示核矩阵,G_i 表示 $\boldsymbol{G}$ 的第 i 列,式(8)可以表示为:

$$\min F(\alpha) = -\frac{\lambda}{n}\sum_{i=1}^{n} y_i\boldsymbol{\alpha}^{\mathrm{T}}G_i + \frac{1}{n}\sum_{i=1}^{n}\max\left(0, \frac{\mathrm{e}^{2(1-y_i\boldsymbol{\alpha}^{\mathrm{T}}G_i)}-1}{\mathrm{e}^{2(1-y_i\boldsymbol{\alpha}^{\mathrm{T}}G_i)}+1}\right) \quad (11)$$

对于变形后的优化问题式(11),不再将其转换成对偶形式,而是使用提出的非线性 MA-ZOVR 算法直接求解。非线性 MA-ZOVR 算法和 2.1 节提到的线性算法有着相同的框架,不同是非线性 MA-ZOVR 算法优化变量为 α 引入了核运算。

非线性 MA-ZOVR 算法如算法 3 所示。

算法 3　非线性 MA-ZOVR 算法

输入　外层迭代轮数 S,内层迭代次数 T,光滑参数 μ,正则化参数 λ

输出　$\hat{\alpha}_S$

初始化　$v_0 = \tilde{\alpha}_0$,$\{\rho_s\}$,$\beta > 0$,η_0

1. for s = 1,2,⋯,S
2. $\tilde{\alpha} = \tilde{\alpha}_{s-1}$
3. 计算全梯度 $\hat{\nabla} F(\tilde{\alpha})$
4. $\eta_s = \dfrac{\eta_0}{\max\left(\beta, \dfrac{2}{s+1}\right)}$
5. $\alpha_0 = \rho_s v_0 + (1-\rho_s)\tilde{\alpha}$
6. for t = 0,1,⋯,T − 1
7. 随机抽取一个样本 i,进行梯度更新
8. $v_{t+1} = v_t - \eta_s(\hat{\nabla} F_i(\alpha_t) - \hat{\nabla} F_i(\tilde{\alpha}) + \hat{\nabla} F(\tilde{\alpha}))$
9. $\alpha_{t+1} = \tilde{\alpha} + \rho_s(v_{t+1} - \tilde{\alpha})$
10. end for
11. $\tilde{\alpha}_s = \dfrac{1}{T}\sum_{t=1}^{T}\alpha_t$
12. $v_0 = v_T$
13. end for
14. $\hat{\alpha}_S = \tilde{\alpha}_S$, if $F(\tilde{\alpha}_S) \leqslant F\left(\dfrac{1}{S}\sum_{s=1}^{S}\tilde{\alpha}_s\right)$,

$\hat{\alpha}_S = \dfrac{1}{S}\sum_{s=1}^{S}\tilde{\alpha}_s$, otherwise

2.3 MA-ZOVR 算法的收敛性分析

在给出 MA-ZOVR 算法的收敛性结论前，对 MA-ZOVR 算法用到的相关知识进行总结如下：

1）整体框架：使用零阶减小方差优化框架，降低了方差的影响，同时避免了重复的梯度计算。

2）梯度计算方法：使用坐标梯度估计[12,15]。虽然多了 $O(d)$ 次的函数查询（计算函数值），但是可以获得更精确的梯度估计。

3）梯度更新方法：在减小方差的基础上结合动量加速技巧[18,21-22]，加速算法的收敛。

在考虑零阶优化的前提下，根据文献[15]关于坐标梯度估计的收敛性分析（定理 3）以及文献[18]中关于动量加速技巧的收敛性分析（定理 1），可以得出本文提出的 MA-ZOVR 算法的收敛速度为 $O\left(\frac{d}{T^2}\right)$。其中，$T=S\times m$ 表示总的迭代次数，d 表示样本维数。同时，根据文献[8]给出的 Pegasos 算法的收敛速率 $O\left(\frac{\log_a T}{T}\right)$ 以及文献[15]给出的 ZO-SVRG 算法的收敛速率 $O\left(\frac{d}{T}\right)$，可以得出本文提出的 MA-ZOVR 算法的收敛速率优于上述 2 个算法。

3 实验结果与分析

本节验证本文提出的动量加速零阶减小方差的鲁棒 SVM（MA-ZOVR）算法的性能，主要从以下 5 个方面进行实验：

1）抗噪性实验：进行线性 MA-ZOVR 算法和非线性 MA-ZOVR 算法的抗噪性实验。

2）模块化实验：分别验证新的求解方法和新的优化模型的有效性。

3）方差分析实验：针对非线性情况，对比本文提出的算法和标准随机梯度下降算法（SGD），验证算法减小方差策略的有效性。

4）收敛速度分析实验：针对非线性情况，对比本文的动量加速零阶减小方差算法和原始的零阶减小方差（ZO-SVRG）算法，验证算法收敛速度的加速。

5）参数分析实验：分析实验中的主要参数对算法精度的影响。

实验程序运行环境为 Matlab R2016a。实验数据来源于 KEEL 网站（训练集和测试集的比例均为 4∶1），主要分为 2 个部分：第 1 部分为常规噪声数据集，依次选取含有 0%、5%、10% 和 15% 属性噪声的数据进行实验；第 2 部分为相对较大规模的标准数据集，为了验证算法的抗噪性，依次对这几个数据集加入 0%、5%、10% 和 15% 均值为 0 及方差为 0.5 的高斯噪声。实验数据集如表 1 所示。

表 1 实验数据集

Table 1 Experimental datasets

数据集	样例数	特征数
sonar	208	60
ionosphere	351	33
wdbc	569	30
pima	768	8
tic-tac-toe	958	9
titanic	2 201	3
penbased	10 992	16
magic	19 020	10
shuttle	58 000	9

3.1 抗噪性实验

本节进行线性 MA-ZOVR 算法和非线性 MA-ZOVR 算法的抗噪性实验。为了使实验效果更加明显，下面给出本文提出的 MA-ZOVR 算法和传统的 SVM 算法（文献[8]中的 Pegasos 算法求解优化问题 1 以及原始零阶减小方差算法，文献[15]中的 ZO-SVRG 算法求解优化问题 1）的测试分类结果。下文所给结果均为五折交叉实验的平均值。线性算法的比较结果如表 2 所示。

表 2 不同线性算法分类准确率比较

Table 2 Comparison of classification accuracy of different linear algorithms %

数据集	0% 噪声			5% 噪声			10% 噪声			15% 噪声		
	MA-ZOVR	Pegasos	ZO-SVRG	MA-ZOVR	Pegasos	ZO-SVRG	MA-ZOVR	Pegasos	ZO-SVRG	MA-ZOVR	Pegasos	ZO-SVRG
sonar	80.95	71.43	76.19	78.57	69.05	73.81	76.19	66.67	71.43	73.81	64.29	66.67
ionosphere	78.87	71.83	76.06	78.87	69.01	74.65	73.24	67.61	71.83	72.28	66.20	67.61
wdbc	85.96	80.70	85.09	79.82	76.32	78.07	78.95	73.68	74.56	72.81	69.30	67.54
pima	66.88	65.58	64.94	65.58	63.64	64.29	64.29	62.99	62.99	63.64	61.69	61.04
tic-tac-toe	66.15	63.54	64.58	65.63	62.50	63.02	64.58	60.94	61.98	64.06	59.38	59.90
titanic	75.28	70.75	71.66	74.60	70.29	70.52	73.70	69.16	70.29	72.56	65.08	69.84
penbased	80.99	75.76	76.85	80.85	73.94	76.49	78.31	71.90	75.44	76.63	68.89	74.62
magic	72.13	69.98	69.06	71.14	68.77	66.61	70.56	65.09	65.96	69.93	63.80	65.14
shuttle	98.20	93.02	95.10	94.17	92.99	94.06	94.09	91.89	94.00	94.04	88.97	92.29

从表 2 可以看出,在给定的 9 组不同噪声比的数据集中,本文提出的线性 MA-ZOVR 算法与线性 Pegasos 算法以及线性 ZO-SVRG 算法相比,均具有较高的准确率,这说明了本文提出的算法有效地提高了 SVM 的抗噪性,具有较高的分类精度。另外,从给出的结果还可以看出,随着数据噪声百分比的增大,分类准确率随之降低。

由于非线性 MA-ZOVR 算法涉及到了核矩阵的运算,因此在处理较大规模数据时,运行时间过长。为了提高实验效率,在常规数据集上进行非线性 MA-ZOVR 算法的抗噪性对比实验。非线性 MA-ZOVR 算法、非线性 Pegasos 算法以及非线性 ZO-SVRG 算法的实验结果如表 3 所示,其中均使用高斯核函数。

表 3　不同非线性算法分类准确率比较

Table 3　Comparison of classification accuracy of different nonlinear algorithms　%

数据集	0% 噪声			5% 噪声			10% 噪声			15% 噪声		
	MA-ZOVR	Pegasos	ZO-SVRG	MA-ZOVR	Pegasos	ZO-SVRG	MA-ZOVR	Pegasos	ZO-SVRG	MA-ZOVR	Pegasos	ZO-SVRG
sonar	76.19	73.80	73.81	74.38	69.04	71.43	73.81	66.67	71.43	66.67	61.90	59.52
ionosphere	80.28	76.05	78.87	77.46	71.83	74.46	77.46	70.42	69.01	76.06	67.71	69.01
wdbc	89.47	85.96	87.72	89.35	82.45	86.84	85.96	78.94	85.09	82.46	71.92	81.58
pima	77.27	67.53	70.13	76.62	65.58	67.53	74.03	62.39	65.08	74.03	60.12	63.18
tic-tac-toe	72.92	67.70	68.23	72.40	65.62	68.23	72.40	63.02	67.71	70.83	61.46	66.67
titanic	78.68	75.28	75.74	78.68	74.37	73.47	75.74	73.24	71.88	75.28	72.33	68.25

从表 3 可以看出,在给定的不同噪声比的 6 组数据集中,本文提出的非线性 MA-ZOVR 算法与非线性 Pegasos 算法以及非线性 ZO-SVRG 算法相比,均具有较高的准确率,这说明非线性 MA-ZOVR 算法的抗噪性能好,能够改善传统 SVM 的不足,降低了噪声数据对分类效果的影响。

3.2　模块化实验

本文提出的 MA-ZOVR 算法包括对优化模型和求解方法 2 个方面的改进。为更好地说明问题,下文进行模块化实验。模块化实验 1:对改进后的优化模型式(7)、式(10)分别使用传统的随机梯度下降的求解方法,记为 MA + SGD;模块化实验 2:对传统的 SVM 模型式(1)使用提出的 MA-ZOVR 优化求解方法(包括线性和非线性),记为 SVM + MA。线性算法模块化实验结果如表 4 所示。

表 4　不同线性算法模块化实验结果比较

Table 4　Comparison of modularization experimental results of different linear algorithms　%

数据集	0% 噪声			5% 噪声			10% 噪声			15% 噪声		
	MA-ZOVR	MA + SGD	SVM + MA	MA-ZOVR	MA + SGD	SVM + MA	MA-ZOVR	MA + SGD	SVM + MA	MA-ZOVR	MA + SGD	SVM + MA
sonar	80.95	78.57	76.19	78.57	76.19	76.19	76.19	72.42	73.81	73.81	71.43	71.43
ionosphere	78.87	76.06	76.06	78.87	74.65	76.06	73.24	72.24	70.42	72.28	71.83	70.42
wdbc	85.96	82.46	84.21	79.82	77.19	76.82	78.95	76.32	75.44	72.81	70.18	67.54
pima	66.88	65.58	65.53	65.58	64.94	63.64	64.29	63.64	63.64	63.64	62.99	60.39
tic-tac-toe	66.15	65.63	65.10	65.63	63.54	63.54	64.58	61.46	61.98	64.06	62.42	61.46
titanic	75.28	74.94	74.28	74.60	72.34	72.11	73.70	71.20	71.43	72.56	70.98	70.29
penbased	80.99	79.40	79.72	80.85	76.72	75.44	78.31	73.99	74.58	76.63	71.90	73.08
magic	72.13	71.45	70.74	71.14	69.77	69.79	70.56	68.51	68.98	69.93	67.74	67.90
shuttle	98.20	97.13	98.02	94.17	94.03	94.04	94.09	93.66	93.98	94.04	93.01	93.05

从表 4 可以看出,在给定的不同噪声比的 9 组数据集中,本文提出的线性 MA-ZOVR 算法的精度优于实验 1 的精度,这说明了在线性情况下本文提出的求解方法的有效性。同样,根据与实验 2 结果的对比也可以看出,在线性情况下本文提出的优化模型的有效性。非线性 MA-ZOVR 算法的模块化实验结果如表 5 所示。从表 5 可以看出,在给定的不同噪声比的 6 组数据集中,本文提出的非线性MA-ZOVR 算法分类精度均高于实验 1和实验 2。按照上述的分析方法,分别验证了在非线性情况下本文提出的求解方法和优化模型的有效性。

表 5 不同非线性算法模块化实验结果比较

Table 5 Comparison of modularization experimental results of different nonlinear algorithms %

数据集	0% 噪声			5% 噪声			10% 噪声			15% 噪声		
	MA-ZOVR	MA + SGD	SVM + MA	MA-ZOVR	MA + SGD	SVM + MA	MA-ZOVR	MA + SGD	SVM + MA	MA-ZOVR	MA + SGD	SVM + MA
sonar	76.19	75.19	74.08	74.38	72.62	71.43	73.81	70.42	71.43	66.67	61.90	59.52
ionosphere	80.28	78.87	78.87	77.46	76.06	76.03	77.46	74.08	74.65	76.06	73.24	70.42
wdbc	89.47	88.60	87.72	89.35	84.72	86.84	85.96	84.21	85.09	82.46	80.70	81.58
pima	77.27	76.62	65.58	76.62	73.38	65.58	74.03	72.08	63.64	74.03	66.88	63.64
tic-tac-toe	72.92	71.88	66.15	72.40	71.35	66.15	72.40	70.83	65.63	70.83	69.79	63.02
titanic	78.68	77.55	74.60	78.68	75.51	68.53	75.74	73.70	68.48	75.28	71.88	67.57

3.3 方差分析实验

本节进行非线性 MA-ZOVR 算法和传统随机梯度下降(Stochastic Gradient Descent,SGD)算法的方差分析实验。图 2 和图 3 为 MA-ZOVR 算法和传统 SGD 算法对不同噪声比的 wdbc 数据集进行分类的方差对比。

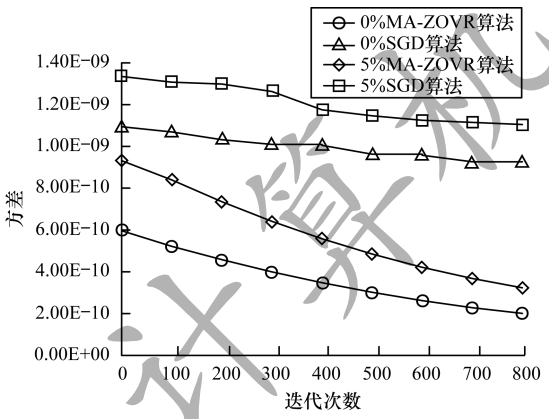

图 2 不同噪声比(0%,5%)的方差对比结果

Fig. 2 Variance comparison results of different noise ratios (0%,5%)

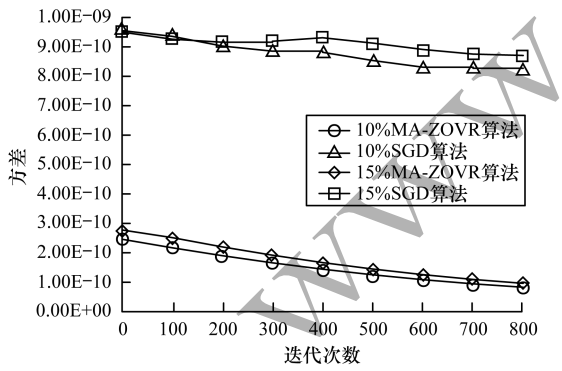

图 3 不同噪声比(10%,15%)的方差对比结果

Fig. 3 Variance comparison results of different noise ratios (10%,15%)

从图 2 和图 3 可以看出,在迭代过程中对于不同噪声比的实验数据,随机梯度下降算法的方差一直维持一个比较大的值,下降幅度较小。对比结果可以看出,本文提出的 MA-ZOVR 算法的方差较小,且随着迭代的进行逐步降低,最后减小到一个接近于零的定值,表明本文算法可以有效地对方差进行修正。

3.4 收敛速度分析实验

本节进行非线性 MA-ZOVR 算法和非线性 ZO-SVRG 算法的收敛速度对比实验。图 4 和图 5 为 2 种方法对不同噪声比的 ionosphere 数据进行分类的收敛速度对比。

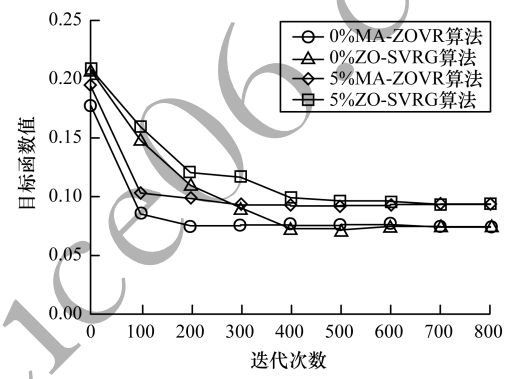

图 4 不同噪声比(0%,5%)的目标函数值

Fig. 4 Objective function values of different noise ratios (0%,5%)

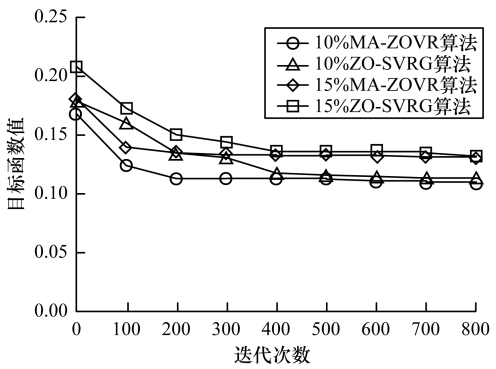

图 5 不同噪声比(10%,15%)的目标函数值

Fig. 5 Objective function values of different noise ratios (10%,15%)

从图 4 和图 5 可以看出,对于不同百分比的噪声数据,使用本文提出的 MA-ZOVR 算法进行求解时,前 200 次迭代过程中函数值逐步减小,在 200 次

以后函数值逐渐趋近于一个定值;对比结果可以看出,原始的 ZO-SVRG 算法在前 400 次迭代中函数值一直处于减小状态,直到 400 次之后函数值才逐渐趋于稳定,表明本文算法通过引入动量加速技巧,有效地提高了算法的收敛速度。

3.5　参数分析实验

本节进行线性 MA-ZOVR 算法和非线性 MA-ZOVR 算法的参数分析实验。对于线性 MA-ZOVR 算法主要分析正则化参数 λ 对分类精度的影响;对于非线性 MA-ZOVR 算法分析正则化参数 λ 和高斯核函数的宽度 sigma 两者共同对分类精度的影响。表 6、表 7 给出不同噪声比的 ionosphere 数据分类的准确率。

表 6　线性 MA-ZOVR 算法分类准确率

Table 6　Classification accuracy of linear MA-ZOVR algorithm　%

噪声	分类准确率		
	$\lambda=0.001$	$\lambda=0.01$	$\lambda=0.1$
0	78.87	66.20	66.20
5	78.87	64.79	66.20
10	73.24	69.01	67.61
15	72.28	59.15	65.63

表 7　非线性 MA-ZOVR 算法分类准确率

Table 7　Classification accuracy of nonlinear algorithm MA-ZOVR

噪声/%	sigma	分类准确率/%		
		$\lambda=0.001$	$\lambda=0.01$	$\lambda=0.1$
0	1	80.28	78.87	77.46
	3	78.87	77.46	77.08
	5	63.38	61.97	62.68
5	1	77.46	77.08	76.98
	3	76.06	75.38	75.32
	5	64.79	63.38	64.08
10	1	77.46	76.59	77.02
	3	73.24	72.65	73.01
	5	66.2	65.98	64.32
15	1	76.06	75.84	75.96
	3	66.2	65.52	65.38
	5	64.79	64.39	64.03

对表 6 进行横向对比,在固定数据噪声比的条件下,根据分类精度可以看出,当参数 $\lambda=0.001$ 时的分类效果较好。

首先对表 7 进行横向对比,在固定噪声比和 sigma 的条件下,根据分类精度可以看出,对于给定的不同 λ 值,分类效果差异不大,当 $\lambda=0.001$ 时分类效果稍好于其他的取值;其次对表 7 进行纵向对比,在固定噪声比和 $\lambda=0.001$ 的条件下,根据结果可得,当 sigma = 1 时分类效果较好。因此,当参数 sigma = 1,$\lambda=0.001$ 时分类性能最优。

4　结束语

为提高 SVM 的抗噪性,本文提出一种基于动量加速零阶减小方差的鲁棒 SVM 算法。通过引入间隔均值项和指数形式的损失函数建立新的优化模型,并在零阶减小方差的基础上引入动量加速技术求解优化模型。实验结果表明,该方法能够有效提高 SVM 的抗噪性,降低在迭代中累积的方差,同时加快算法的收敛速度。下一步将在本文研究的基础上,结合 L_1 正则化项,设计新的算法对带有噪声数据分类问题的稀疏化进行研究。

参考文献

[1] VAPNIK V N. The nature of statistical learning theory[M]. Berlin, Germany: Springer, 1995.

[2] GAO Wei, ZHOU Zhihua. On the doubt about margin explanation of boosting[J]. Artificial Intelligence, 2013, 203(2): 1-18.

[3] ZHANG Teng, ZHOU Zhihua. Large margin distribution machine[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2014: 313-322.

[4] WU Chao, LIU Yufeng. Robust truncated hinge loss support vector machines[J]. Journal of the American Statistical Association, 2007, 102(479): 974-983.

[5] HUANG Xiaolin, SHI Lei. Support vector machine classifier with pinball loss[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 984-997.

[6] YANG Liming, DONG Hongwei. Support vector machine with truncated pinball loss and its application in pattern recognition[J]. Chemometrics and Intelligent Laboratory Systems, 2018, 177(2): 89-99.

[7] SHALEV-SHWARTZ S, SINGER Y. PEGASOS: primal estimated sub-gradient solver for SVM[J]. Mathematical Programming, 2011, 127(1): 3-30.

[8] JOHNSON R, ZHANG T. Accelerating stochastic gradient descent using predictive variance reduction[J]. News in Physiological Sciences, 2013, 1(3): 315-323.

[9] NESTEROV Y. Introductory lectures on convex optimization: a basic course[M]. Boston, USA: Kluwer Academic Public, 2004.

[10] SHANG Fanhua, LIU Yuanyuan. Fast stochastic variance reduced gradient method with momentum acceleration for machine learning[EB/OL]. [2019-09-10]. https://www.researchgate.net/publication.

[11] TSENG P. Approximation accuracy, gradient methods, and error bound for structured convex optimization[J]. Mathematical Programming, 2010, 125(2): 263-295.

[12] ZHU Zeyuan. Katyusha: the first direct acceleration of stochastic gradient methods[C]//Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. New York, USA: ACM Press, 2017: 1200-1205.

(下转第 104 页)

(上接第 95 页)

[13] NGUYEN V C, XU Huan. Accelerated stochastic mirror descent algorithms for composite non-strongly convex optimization[J]. Journal of Optimization Theory and Applications, 2019, 18(2): 541-566.

[14] LIU S, KAILKHURA B. Zeroth-order stochastic variance reduction for nonconvex optimization[C]// Proceedings of the 32nd Conference on Neural Information Processing Systems. Washington D. C., USA: IEEE Press, 2018: 1-26.

[15] LIU Shijia, CHEN Jie. Zeroth-order online alternating direction method of multipliers: convergence analysis and applications[C]//Proceedings of the 21st International Conference on Artificial Intelligence and Statistics. Washington D. C., USA: IEEE Press, 2018: 288-297.

[16] ZHU Zeyuan. Katyusha: accelerated variance reduction for faster SGD[EB/OL]. [2019-09-10]. https://arxiv.org/abs/1603.05953v1.

[17] SHANG Fanhua, ZHOU Kaiwen. VR-SGD: a simple stochastic variance reduction method for machine learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(1): 188-202.

[18] CHEN Fanyong, ZHAN Jing. Large cost-sensitive margin distribution machine for imbalanced data classification[J]. Neurocomputing, 2017, 224(8): 45-57.

[19] ZHOU Yuhang, ZHOU Zhihua. Cost-sentive large margin distribution machine[J]. Journal of Computer Research and Development, 2016, 53(9): 1964-1970. (in Chinese)
周宇航, 周志华. 代价敏感大间隔分布学习机[J]. 计算机研究与发展, 2016, 53(9): 1964-1970.

[20] TAO Wei, PAN Zhisong. The individual convergence of projected subgradient methods using the Nesterov's step-size stratey[J]. Chinese Journal of Computers, 2018, 41(1): 164-176. (in Chinese)
陶蔚, 潘志松. 使用 Nesterov 步长策略投影次梯度方法的个体收敛性[J]. 计算机学报, 2018, 41(1): 164-176.

[21] CHEN Yujia, TAO Wei. Optimal individual convergence rate of the Heavy-Ball-Based momentum methods[J]. Journal of Computer Research and Development, 2019, 56(8): 1686-1694. (in Chinese)
程禹嘉, 陶蔚. Heavy-Ball 型动量方法的最优个体收敛速率[J]. 计算机研究与发展, 2019, 56(8): 1686-1694.

编辑　索书志