一种基于 Comid 的非光滑损失随机坐标下降方法

陶 卿,朱烨雷,罗 强,孔 康
(中国人民解放军陆军军官学院,安徽合肥 230031)

摘 要: 坐标下降方法以简洁的操作流程、低廉的计算代价和快速的实际收敛效果,成为处理大规模优化最有效的方法之一.但目前几乎所有的坐标下降方法都由于子问题解析求解的需要而假设损失函数的光滑性.本文在结构学习的框架下,在采用 Comid 方法求解随机挑选单变量子问题的基础上,提出了一种新的关于非光滑损失的随机坐标下降方法.理论分析表明本文所提出的算法在一般凸条件下可以得到 $O(\sqrt{t}/t)$的收敛速度,在强凸条件下可以得到 $O(\ln t/t)$的收敛速度.实验结果表明本文所提出的算法对正则化 Hinge 损失问题实现了坐标优化预期的效果.

关键词: 机器学习; 优化; 大规模; 坐标下降方法; 非光滑损失; 结构学习; COMID.
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2013)04-0768-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.04.024

A New Comid-Based Stochastic Coordinate Descent Method for Non-Smooth Losses

TAO Qing, ZHU Ye-lei, LUO Qiang, KONG Kang
(*Chinese People's Liberation Army Officer Academy, Hefei, Anhui* 230031, *China*)

Abstract: Coordinate descent (CD) method is one of the most efficient algorithms in dealing with the large-scale optimization problems for its simple operation, cheap computational cost and practical efficiency. However until now, almost all the-state-of-art CD algorithms require the smoothness assumption of loss functions due to solving the subproblems in closed-form. In this paper, within the structural learning framework, we present a new stochastic CD(SCD) algorithm for non-smooth losses, in which the randomly selected single variable problem is solved using Comid method. Theoretical analysis indicates that the proposed algorithm has an $O(\sqrt{t}/t)$ convergence rate for general convex problems and an $O(\ln t/t)$ convergence rate for strongly convex problems. The experiments demonstrate the expected efficiency of our SCD algorithms when coping with the L1-regularized Hinge loss problems.

Key words: machine learning; optimization; large-scale; coordinate descent methods; non-smooth losses; structural learning; COMID

1 引言

统计机器学习无论是在理论上还是在应用上都取得了极为丰硕的成果.当前,一般在"正则化项 + 损失函数"的框架下对学习问题进行理论分析,这是对基于"间隔"的统计机器学习理论框架的进一步发展和完善[1].随着互联网的日益普及和计算机技术的迅猛发展,机器学习所面临的问题逐渐呈现出样本多、维数高的特点,一个普通文本数据库就会达到 10^7 样本个数或 10^9 样本维数的规模.坐标下降方法(Coordinate Descent, CD)能够利用高维数据的稀疏特性[2],是解决大规模高维数据问题卓有成效的手段之一.大量的比较实验表明 CD 是解决大规模文本分类问题的首选算法[3~5].

Paul Tseng 是坐标优化算法研究方面的著名学者,他从优化理论的角度对满足光滑性和凸性假设目标函数的优化问题建立了坐标优化算法并讨论了收敛性,其中也包括了机器学习正则化光滑损失函数的原优化问题和正则化非光滑损失的对偶优化问题[6,7].原始问题的坐标下降方法(Primal Coordinate Descent, PCD)是坐标下降方法中一种最常见形式,其主要思路是通过逐个优化解向量 $\boldsymbol{W}$ 的每一维特征(坐标),实现一次外循环.内循环中,在优化某一坐标时,固定解向量中的其余 n - 1

收稿日期:2012-02-09;修回日期:2012-10-24
基金项目:国家自然科学基金项目(No.61273296, No.60975040)

维坐标不动,一次仅对选中的一维坐标求解单变量子问题.PCD 具有简洁操作流程,已经成功应用到信号处理[8]、文本分类、图像恢复等诸多领域[2,3,9,10].CD 的成功应用主要得益于计算代价低廉的方向导数计算技巧[2]以及各维特征的实时更新[11].著名学者 Nesterov 甚至认为 CD 如果没有计算代价低廉的方向导数计算技巧,CD 就失去了生存的依据[2].

单变量优化子问题的求解是各种坐标下降方法的核心,不同坐标优化算法的区别也主要体现在单变量问题的求解方式上.正是由于单维子问题解析求解的需要,目前几乎所有的坐标下降方法都假设了损失函数的光滑性(导数 Lipschitz 连续),或者将原问题转化为等价的对偶问题使目标具有光滑性[4],如文献[2,3,9,10]中不同 PCD 算法的共同点是在光滑性的前提下对损失函数进行二阶展开[12].但另一方面,机器学习中普遍使用的支持向量机算法所采用的 Hinge 损失[4,13]却是非光滑的,由于此时子问题的解析求解存在着一定的困难,已有坐标优化方法的设计思路难以直接平移得到求解支持向量机的 PCD,这也是非光滑损失的优化问题至今仍然没有坐标优化方法的主要原因.

尽管支持向量机的损失函数是非光滑的,但其对偶问题的目标函数却是二次多项式形式的,虽然存在着约束条件,对应的子问题还是可以很方便地解析求解.正是利用这一事实,林智仁教授领导的研究小组于 2008 年提出了对偶坐标下降方法(Dual Coordinate Descent,DCD)求解支持向量机的对偶问题[4].在大量的数据库上,DCD 取得了比 PCD 和当前流行的一些算法更快的收敛效果.然而,DCD 并不能完全取代 PCD.正如文献[3]所指出的那样,PCD 和 DCD 各有长处.一般地说,相比于 DCD,PCD 更适合一些样本特征维数远小于样本个数的学习问题.另外,很多非光滑损失的原优化问题如"L1 正则化 + Hinge 损失"也不存在对偶形式.因此,对非光滑损失函数的优化问题如何建立 PCD 无论在理论还是在应用上都是一个亟待解决的问题.

近些年来,凸优化问题结构方法的研究取得了一批具有深远影响的结果.著名学者 Nesterov 曾经断言:"传统的黑箱(Black-Box)方法在凸优化问题上的重要性将不可逆转地消失,彻底地取而代之的是巧妙运用问题结构的新算法"[14].对于非光滑目标函数的优化问题,Nesterov 于 2009 年提出了一种原始-对偶(primal-dual)次梯度方法[15]代替传统的投影次梯度算法(Projected Sub-gradient Algorithm, PSA)[5].而对光滑的目标函数,Nesterov 的加速方法达到了优于一般"黑箱"方法的一个数量级的结论[15].2009 年,Xiao 将 Nesterov 在文献[14,15]中的工作拓展为可以处理正则化形式机器学习问题的在线算法,得到了正则化共轭平均(Regularized Dual Averaging,RDA)[16].这种方法克服了在线 PSA 算法不能保证正则化项结构的缺陷,特别对于 L1 正则化问题可以获得和批处理算法同等程度的稀疏解.2010 年,Duchi 等将著名的镜面下降(Mirror Descent,MD)方法进行了拓展,得到了复合目标函数的镜面下降算法(Composite Objective Mirror Descent,COMID)[17],统一了前期关于在线算法的很多研究.与 RDA 方法类似,Comid 不仅能有效保证正则化项的结构,而且无论是对光滑还是非光滑损失均可应用.不同的是,Comid 以瞬时梯度替代了 RDA 的平均梯度,更加简洁和易操作.本文的主要思路是在采用 Comid 方法求解随机挑选单变量子问题的基础上,提出了一种新的非光滑损失的随机坐标下降方法 Non-smooth SCD.理论分析表明 Non-smooth SCD 在一般凸条件下可以得到 $O(\sqrt{t}/t)$的收敛速度,在强凸条件下可以得到 $O(\ln t/t)$的收敛速度.实验结果表明,Non-smooth SCD 优于当前流行的如文献[4,16~19]中的一些算法,达到了随机坐标优化预期的效果.

2　正则化学习问题与坐标优化方法

与所有坐标优化算法的文献一样,为简单明起见,本文仅讨论线性无偏置项的二分类问题.设 $S=\{(\boldsymbol{X}_1,y_1),\cdots,(\boldsymbol{X}_m,y_m)\}\in\mathbf{R}^n\times\{+1,-1\}$是一些独立同分布样本组成的训练集,正则化机器学习框架可以归结为求解下述凸优化问题:

$$\min_{\boldsymbol{W}} F(\boldsymbol{W})=\sum_{i=1}^{m} f_i(\boldsymbol{W})+P(\boldsymbol{W}) \tag{1}$$

其中 $\boldsymbol{W}\in\mathbf{R}^n$,损失函数 $f_i(\boldsymbol{W})$表示由样本 $\boldsymbol{X}_i$ 造成的损失.本文主要研究 $P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_1$和 $P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_2$下的 Hinge 损失问题.$f_i(\boldsymbol{W})=\max\{0,1-y_i(\langle\boldsymbol{W},\boldsymbol{X}_i\rangle)\}$有时也称为 L1 损失.由于 Hinge 损失仅仅是次可微的,因此是非光滑的.$f_i(\boldsymbol{W})=\max\{0,1-y_i(\langle\boldsymbol{W},\boldsymbol{X}_i\rangle)\}^2$通常称为 L2 损失,其导数 Lipschitz 连续因而具有光滑性.

坐标下降方法的迭代过程分为内循环和外循环两部分[3].迭代过程从起始点 $\boldsymbol{W}^0$ 开始依次迭代出 $\boldsymbol{W}^1,\boldsymbol{W}^2,\cdots,\boldsymbol{W}^T$.从 $\boldsymbol{W}^t$ 到 $\boldsymbol{W}^{t+1}$的过程称为一次外循环.$\boldsymbol{W}^{t+1}$通过更新 $\boldsymbol{W}^t$ 的 n 个变量来实现一次外循环,每一次外循环包含 n 次内循环.每次内循环生成 $\boldsymbol{W}^{t,j}\in\mathbf{R}^n,j=1,\cdots,n$,并且 $\boldsymbol{W}^{t,1}=\boldsymbol{W}^t$,$\boldsymbol{W}^{t,n+1}=\boldsymbol{W}^{t+1}$,$\boldsymbol{W}^{t,j}=[w_1^{t+1},\cdots,w_{j-1}^{t+1},w_j^t,\cdots,w_n^t]^{\mathrm{T}},j=2,\cdots,n$.对于 $\boldsymbol{W}^{t,i}$到 $\boldsymbol{W}^{t,j+1}$的更新,通过求解如下单变量子问题得到:

$$\begin{aligned}\min_z \boldsymbol{W}^{t,j}&=[w_1^{t+1},\cdots,w_{j-1}^{t+1},w_j^t+z,w_{j+1}^t,\cdots,w_n^t]^{\mathrm{T}}\\&=\min_z f(\boldsymbol{W}^{t,j}+z\boldsymbol{e}_j)\end{aligned} \tag{2}$$

其中 $\boldsymbol{e}_j=[0,\cdots,1,\cdots 0]^{\mathrm{T}}$.

坐标下降方法的算法流程见下面的算法 1:

算法 1 坐标下降方法

·Start with any initial $\boldsymbol{W}^0$

·For $t=0,1,\cdots,T$ (outer iterations)

For $j=1,2,\cdots,n$ (inner iterations)

Fix $w_1^{t+1},\cdots,w_{j-1}^{t+1},w_{j+1}^t,\cdots,w_n^t$ and approximately solve the sub-problem(2)

式(1)的单变量优化问题可以写成下面形式,

$$\min_z \quad a_j(z)+P(\boldsymbol{W}^{t,j}+z\boldsymbol{e}_j) \tag{3}$$

其中 $a_j(z)=f(\boldsymbol{W}^{t,j}+z\boldsymbol{e}_j)$. 在式(3)中,如果损失函数是光滑的,可以根据光滑性对损失函数进行二阶近似展开,得到如下的优化问题

$$\min_z a_j'(0)z+\frac{1}{2}a_j''(0)z^2+p(w_j^{t,j}+z)-p(w_j^{t,j}) \tag{4}$$

其中 $a_i''(0)$是 $a_i'(0)$在 0 处的次梯度[3,13]. 子问题(4)变成一个关于单变量 z 的二项式,文献[3]针对光滑的 L2 损失,用牛顿法进行近似求解.

对于高维问题,循环坐标方法一次外循环就需要遍历所有的维数,计算和存储代价非常大. 如果每次内循环中只随机抽取一维坐标并对其进行更新,则会大大减少计算代价,这就是随机坐标下降方法(SCD)的主要思路. 本文重点研究 SCD.

3 非光滑损失的随机坐标下降方法

假设 $f(\boldsymbol{W})$是凸且连续的函数,$\partial f(\boldsymbol{W})$表示函数 f 在 $\boldsymbol{W}$ 处的次梯度. 显然,对于非光滑损失函数的情况,无法直接使用式(4)对损失函数二阶近似展开. 本文将在结构学习框架下使用批处理形式的 Comid 算法求解非光滑损失对应的子问题.

3.1 Comid 算法介绍

虽然一阶梯度方法求解最优解的速度稍慢,但仍然能较快地获得稳定的学习精度. 当样本维数足够大时,MD 被认为是最优的一阶方法[20]. 对不含正则化项的优化问题,在线 MD 算法的主要迭代如下:

$$\boldsymbol{W}^{t+1}=\arg\min_{\boldsymbol{W}}\{B_\varphi(\boldsymbol{W},\boldsymbol{W}^t)+\eta_t\langle \boldsymbol{g}^t,\boldsymbol{W}-\boldsymbol{W}^t\rangle\} \tag{5}$$

这里 B_φ 表示凸函数 φ 的 Bregman Divergence, $\boldsymbol{g}^t\in\partial f_t(\boldsymbol{W}^t)$,$\langle\cdot,\cdot\rangle$表示向量间的内积,$\eta_t$ 为步长参数. 对于正则化学习问题,如果将正则化项与损失函数同等对待而直接使用 MD 方法,就会失去正则化项的结构作用,变成了 Nesterov 称之为的"黑箱"方法[14]. 特别当 $P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_1$ 时,直接使用式(5)求解,就会失去解的稀疏性. 与黑箱方法不同的是,结构学习方法只是对损失函数进行近似展开,而保持正则化项不动,此时子问题可以解析求解. 具体来说,在结构学习框架下,在线算法 Comid 的主要步骤为:

$$\boldsymbol{W}^{t+1}=\arg\min_{\boldsymbol{W}}\{\eta_t\langle g^t,W-W^t\rangle+B_\varphi(\boldsymbol{W},\boldsymbol{W}^t)+\eta_tP(\boldsymbol{W})\} \tag{6}$$

不失一般性,在本文的坐标优化算法中,我们取 $B_\varphi(\boldsymbol{W},\boldsymbol{W}^t)=\|\boldsymbol{W}-\boldsymbol{W}^t\|_2^2$,对(6)进一步简化可得,

$$\boldsymbol{W}^{t+1}=\arg\min_{\boldsymbol{W}}\{\langle \eta_t\boldsymbol{g}^t-2\boldsymbol{W}^t,\boldsymbol{W}\rangle+\eta_tP(\boldsymbol{W})+\|\boldsymbol{W}\|_2^2\} \tag{7}$$

显然,式(7)是可分问题,即解向量的各维可以相互独立地进行求解,即只需求解

$$w_j^{t+1}=\arg\min_w\{w^2+\eta_t p(w)+w(\eta_t g_j^t-2w_j^t)\} \tag{8}$$

其解析解的求法见附录 1. 从解析解中,不难看出在线算法 Comid 计算代价低同时又能保证正则化项的结构,这正是结构学习的优点. 基于上述考虑,本文提出的坐标优化算法使用批处理形式的 Comid 算法求解子问题,其具体描述见算法 2,它与在线形式 Comid 的主要区别是此时 $\boldsymbol{g}^t\in\partial f(\boldsymbol{W}^t)$而不是 $\boldsymbol{g}^t\in\partial f_t(\boldsymbol{W}^t)$.

与批处理 Comid 一次迭代遍历所有维不同的是,本文提出的 Non-smooth SCD 仅仅在某一维 j 上求解式(8),这里 j 是从 n 维坐标中等概率随机抽取的. 如果 w_j^{t+1} 是该子问题的解,解向量 $\boldsymbol{W}^t$ 到 $\boldsymbol{W}^{t+1}$的更新为 $\boldsymbol{W}^{t+1}=[w_1^t,\cdots,w_j^{t+1},\cdots w_n^t]$. Non-smooth SCD 算法具体描述见算法 3.

算法 2 批处理 Comid 算法

·Initialize a weight vector $\boldsymbol{W}^0=0$ and $\boldsymbol{g}^0=\boldsymbol{0}$

·repeat

1. compute $\boldsymbol{g}^t\in\partial f(\boldsymbol{W}^t)$
2. compute $\boldsymbol{W}^t$ via(7)
3. $t:=t+1$

until a stopping condition is satisfied

算法 3 非光滑随机坐标下降方法(Non-smooth SCD)

·Initialize a weight vector $\boldsymbol{W}^0=0$ and $\boldsymbol{g}^0=0$

·repeat

1. choose $j\in\{1,2,\cdots,n\}$ uniformly at random
2. let g_j^t be the j-th element of $\boldsymbol{g}^t\in\partial f(\boldsymbol{W}^t)$
3. solve
$$w_j^{t+1}=\arg\min_w\{w^2+\eta_t p(w)+w(\eta_t g_j^t-2w_j^t)\}$$
4. update $\boldsymbol{W}^{t+1}$
5. $t:=t+1$

until a stopping condition is satisfied

值得指出的是,此处更新梯度 $\boldsymbol{g}^t\in\partial f(\boldsymbol{W}^t)$时用到了代价较低的计算技巧. 以 Hinge 损失为例,

$$\boldsymbol{g}^t=\begin{cases}-y_i\boldsymbol{X}_i, & \text{if } y_i\langle\boldsymbol{W},\boldsymbol{X}_i\rangle<1\\ 0, & \text{otherwise}\end{cases}$$

由于求梯度需要判断 $y_i\langle\boldsymbol{W},\boldsymbol{X}_i\rangle$的大小,计算开销主要

来自于内积运算$\langle \boldsymbol{W}, \boldsymbol{X}_i \rangle$.更新一维后,如果重新计算所有样本的$\langle \boldsymbol{W}, \boldsymbol{X}_i \rangle$,其计算复杂度高达 $o(mn)$.坐标下降方法中求方向导数的技巧能够巧妙地避免这一瓶颈,使得更新一次的计算复杂度仅为 $o(m)$.为了说明这种技巧,首先注意到内积矩阵可以写成下面形式:

$$\begin{bmatrix} \langle \boldsymbol{W}, \boldsymbol{X}_1 \rangle \\ \langle \boldsymbol{W}, \boldsymbol{X}_2 \rangle \\ \vdots \\ \langle \boldsymbol{W}, \boldsymbol{X}_m \rangle \end{bmatrix} = \begin{bmatrix} x_{11} \cdots x_{1j} \cdots x_{1n} \\ x_{21} \cdots x_{2j} \cdots x_{2n} \\ \vdots \quad \vdots \quad \vdots \\ x_{m1} \cdots x_{mj} \cdots x_{mn} \end{bmatrix}_{m\times n} \begin{bmatrix} w_1 \\ \vdots \\ w_j \\ \vdots \\ w_n \end{bmatrix}_{n\times 1}$$

$$= \begin{bmatrix} w_1 x_{11} + \cdots + w_j x_{1j} + \cdots + w_n x_{1n} \\ w_1 x_{21} + \cdots + w_j x_{2j} + \cdots + w_n x_{2n} \\ \vdots \quad \vdots \quad \vdots \\ w_1 x_{m1} + \cdots + w_j x_{mj} + \cdots + w_n x_{mn} \end{bmatrix}_{m\times 1}$$

从上面矩阵的分解中不难发现,w_j 对应的是第 j 维所有样本,因此更新 w_j^t 到 w_j^{t+1} 仅仅需要对 j 维的样本进行更新,而其余部分保持不变,如下所示:

$$\begin{bmatrix} \langle \boldsymbol{W}^{t+1}, \boldsymbol{X}_1 \rangle \\ \langle \boldsymbol{W}^{t+1}, \boldsymbol{X}_2 \rangle \\ \vdots \\ \langle \boldsymbol{W}^{t+1}, \boldsymbol{X}_m \rangle \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{W}^{t}, \boldsymbol{X}_1 \rangle \\ \langle \boldsymbol{W}^{t}, \boldsymbol{X}_2 \rangle \\ \vdots \\ \langle \boldsymbol{W}^{t}, \boldsymbol{X}_m \rangle \end{bmatrix} + \begin{bmatrix} (w_j^{t+1} - w_j^t) x_{1j} \\ (w_j^{t+1} - w_j^t) x_{2j} \\ \vdots \\ (w_j^{t+1} - w_j^t) x_{mj} \end{bmatrix}$$

每次更新只需要计算第 j 维$(w_j^{t+1} - w_j^t) x_{ij}$的值,有效地降低了计算复杂度,这就是坐标下降方法收敛速度快的主要原因之一[2].

3.2　随机坐标下降方法收敛性分析

本文基于 Comid 的 SCD 收敛性主要依赖于批处理 Comid 的收敛性.为此,我们引入假设.

假设　存在常数 $G_* > 0$,满足 $\| \boldsymbol{g} \| \leqslant G_*$,其中 $\boldsymbol{g} \in \partial f(\boldsymbol{W})$, $\boldsymbol{W} \in \mathbf{R}^n$.存在常数 $M > 0$,满足 $\| r \| \leqslant M$,其中 $\boldsymbol{r} \in \partial P(\boldsymbol{W})$, $\boldsymbol{W} \in \mathbf{R}^n$.当使用 L1 正则化项时,$\eta_t$ 与$\sqrt{t}$同阶.当使用 L2 正则化项时,$\eta_t = O(1/t)$.

在上述假设下,文献[17]得到了在线算法 Comid 的 regret 界[21,22].基于在线算法和批处理算法之间的关系,可以直接得到:

定理 1　假设$\{\boldsymbol{W}^t\}$由批处理 Comid 产生,对于 $D > 0, \alpha > 0$,令 $F_D = \{B_\varphi(\boldsymbol{W}, \boldsymbol{W}^t) \leqslant D^2\}$, $\boldsymbol{W}^*$ 是 F_D 上算法 2 的最优解且$\frac{\alpha}{2} \| \boldsymbol{W} - \boldsymbol{W}^t \|^2 \leqslant B_\varphi(\boldsymbol{W}, \boldsymbol{W}^t)$.

(1)如果 $P(\boldsymbol{W}) = \lambda \| \boldsymbol{W} \|_1$,则

$$F(\overline{\boldsymbol{W}}^t) - F(\boldsymbol{W}^*) \leqslant O\left(\frac{DG_*^2}{\sqrt{\alpha}} \cdot \frac{\sqrt{t}}{t}\right)$$

(2)如果 $P(\boldsymbol{W}) = \lambda \| \boldsymbol{W} \|_2^2$,则

$$F(\overline{\boldsymbol{W}}^t) - F(\boldsymbol{W}^*) \leqslant O\left(\frac{G_*^2}{\lambda\alpha} \cdot \frac{\ln t}{t}\right)$$

这里 $\overline{\boldsymbol{W}}^t = \left[\sum_{j=1}^{t} \boldsymbol{W}^j\right]/t$.

t 次迭代以后,Non-smooth SCD 算法可以得到$[\boldsymbol{W}^t, F(\boldsymbol{W}^t)]$.这个结果依赖于随机选择的坐标序列集 $\xi_t = \{j_1, j_2, \cdots, j_t\}$,其中 j_t 是在特征集合$\{1, 2, \cdots, n\}$中等概率随机抽取的.这里不妨设 $\phi_t = E_{\xi_{t-1}}[F(\boldsymbol{W}^t)]$.

定理 2　设 $\boldsymbol{W}^t$ 由 Non-smooth SCD 算法产生,$\boldsymbol{W}^*$ 是优化问题的最优解.对于$\forall\ \boldsymbol{W} = (w_1, w_2, \cdots, w_n)^{\mathrm{T}} \in F_D$,定义

$$R_t(\boldsymbol{W}) = \sum_{\tau=1}^{t} \{\phi_\tau - F(\boldsymbol{W})\}$$

$$\delta_t(\boldsymbol{W}) = \sum_{\tau=1}^{t} \{g_{j_\tau}^\tau (w_{j_\tau}^\tau - w_{j_\tau}) + p(w_{j_\tau}^\tau)\} - \frac{1}{n} t P(\boldsymbol{W})$$

则 $R_t(\boldsymbol{W}) \leqslant n E_{\xi_t} \delta_t(\boldsymbol{W})$.

定理 2 证明见附录 2,从定理 2 证明中可以得到:

$$nE_{\xi_t}\delta_t(\boldsymbol{W}) = E_{\xi_t} \sum_{\tau=1}^{t} [g^\tau(\boldsymbol{W}^\tau - \boldsymbol{W}^*) + P(\boldsymbol{W}^\tau) - P(\boldsymbol{W}^*)]$$

与[17]类似,我们可以得到 $\sum_{\tau=1}^{t} [\boldsymbol{g}^\tau(\boldsymbol{W}^\tau - \boldsymbol{W}^*) + \boldsymbol{r}^\tau(\boldsymbol{W}^{\tau+1} - \boldsymbol{W}^*)]$的界,并在此基础上得到如下定理(证明见附录 3):

定理 3　令 $\boldsymbol{W}^*$ 是 F_D 上的最优解,$\{\boldsymbol{W}^t\}$由 Non-smooth SCD 产生,

(1)如果 $P(\boldsymbol{W}) = \lambda \| \boldsymbol{W} \|_1$,则

$$\frac{1}{t} \sum_{\tau=1}^{t} \phi_\tau - F(\boldsymbol{W}^{t*}) \leqslant O\left(\frac{DG_*^2}{\sqrt{\alpha}} \cdot \frac{\sqrt{t}}{t}\right)$$

(2)如果 $P(\boldsymbol{W}) = \lambda \| \boldsymbol{W} \|_2^2$,则

$$\frac{1}{t} \sum_{\tau=1}^{t} \phi_\tau - F(\boldsymbol{W}^*) \leqslant O\left(\frac{G_*^2}{\lambda\alpha} \cdot \frac{\ln t}{t}\right)$$

根据定理 3,我们还能类似于文献[17]得到依概率收敛的速率.

4　实验

本节的主要目的是实验验证 Non-smooth SCD 的实际效果.实验在 Sun Ultra45 工作站(1.6GHz UltraSPARC IIIi 处理器,4GB 内存,Solaris10 操作系统)上实现.本文的 SCD 在 Liblinear 平台上实现.

4.1　大规模数据库描述

本文所采用的四个数据库是 astro-physics、CCAT、a9a 和 covtype,其中 astro-physics 和 CCAT 是文本库.

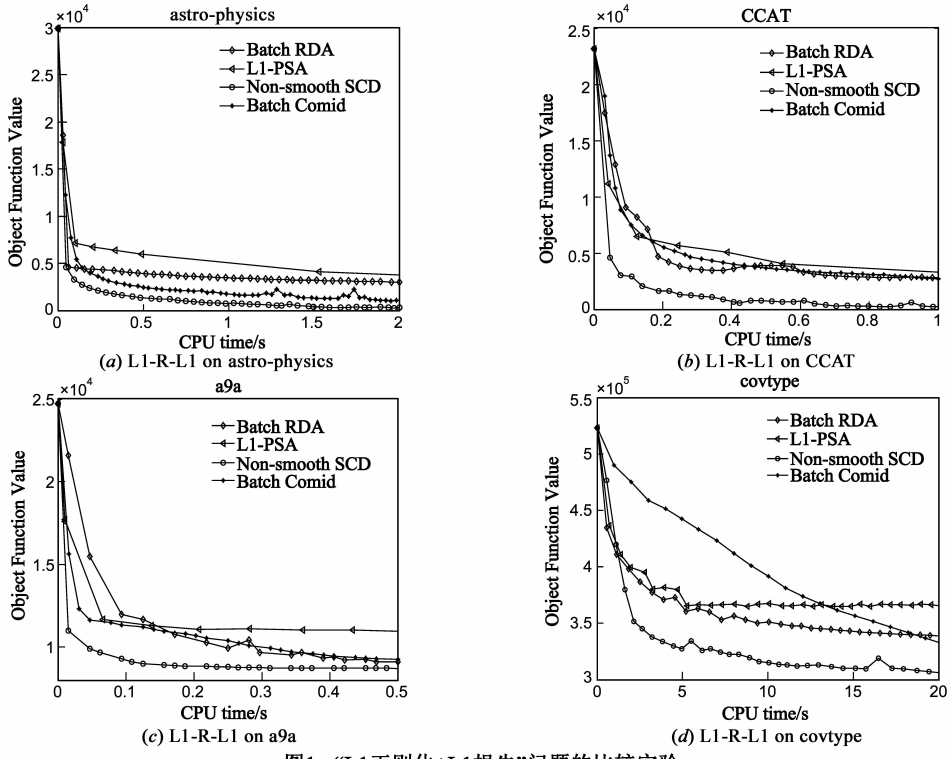
本实验中所有的平衡参数取 $\lambda = 1/m$.为使实验结果更为客观公平,所采用的随机算法均运行 10 次,采取平均后的结果作为标准来进行比较.

表 1 实验数据库描述

数据集	训练样本	测试样本	维数
astro-physics	29,882	32,487	99,757
CCAT	23,149	781,265	47,236
a9a	24,703	7,858	123
covtype	522,911	58,101	54

4.2 “L1 正则化 + Hinge 损失”的实验比较

考虑“L1 正则化 + L1 损失”(L1-R-L1)的情况,由于 L1-R-L1 缺乏强的凸性和光滑性,很少有文献专门对此研究,在综述性文献[13]就没有讨论该问题.我们选择与 L1-PSA[18,19]、Batch RDA[16]以及 Batch Comid[17]进行比较.实验结果如图 1 所示:



(a) L1-R-L1 on astro-physics (b) L1-R-L1 on CCAT

(c) L1-R-L1 on a9a (d) L1-R-L1 on covtype

图1 “L1正则化+L1损失”问题的比较实验

4.3 支持向量机的实验比较

考虑与 L2 正则化项和 L1 损失(L2-R-L1),由于 L2 正则化项具有强凸性质,已有大量文献对此类问题的求解进行了讨论,比较高效的算法包括 Pegasos[23]、DCD[4]、Batch RDA[16]和 Batch Comid[17].值得指出的是,文献[13]中将 DCD 与 SGD[23]、SVM^{perf}[24]、bundle[25]及 TRON[26]做了全面的实验比较,结果表明 DCD 具有更快的实际收敛效果.DCD 主要有循环对偶坐标下降(CD-CD)和随机对偶坐标下降(SDCD)两种形式[4].我们选择与 Pegasos、Batch RDA、Batch Comid、CDCD 以及 SDCD 比较.实验结果如图 2 所示.

4.4 实验结论

基于不同的实验比较对象,发现以下几点现象:

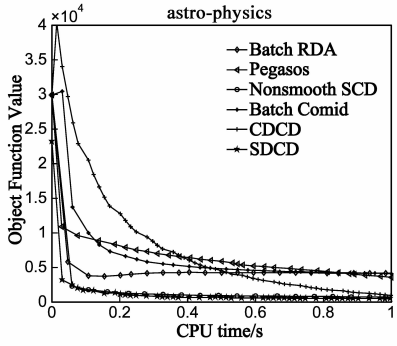
(1)随机与确定性方法的比较:从图 2 中可以看出,随机坐标优化方法 SCD 比循环坐标优化方法 CDCD 收敛得更快.

(2)解原始优化问题算法的比较:从图 1 和 2 中可以看出,Non-smooth SCD 快于 L1-PSA 和 Pegasos.

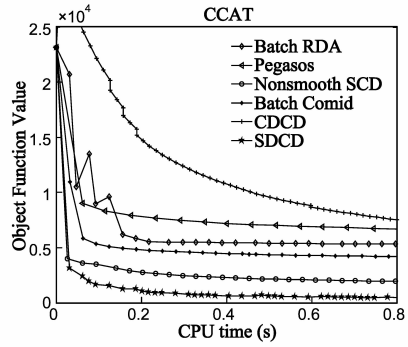
(3)非光滑损失 CD 和批处理算法的比较:从图 1 和 2 中可以看出,Non-smooth SCD 比 Batch Comid 和 Batch RDA 收敛得更快 (稀疏程度几乎相同).

(4)PCD 和 DCD 的比较:当样本个数 m 小于维数 n 时,在 astro-physics 和 CCAT 文本库上,Non-smooth SCD 比 SDCD 稍慢一点,但当 $m \geqslant n$ 时,在 a9a 和 covtype 的非文本库上,Non-smooth SCD 表现出更加优越的性能(图 2a、图 2b).实际上,当样本个数远大于特征数时,文献[3]中的对偶方法甚至会出现在某些数据库上原始问题目标函数不能够快速收敛的现象(图 2(c)和图 2(d)).

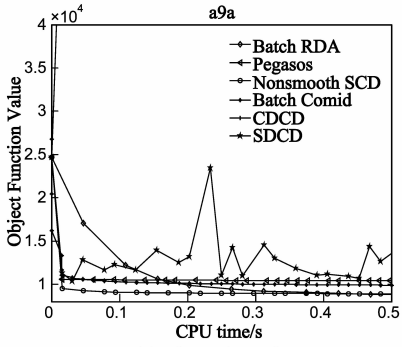
根据上述实验比较,Non-smooth SCD 在求解大规模非光滑原始问题时具有操作简单、计算代价低、保证结构和快速收敛等优点,因此 Non-smooth SCD 实现了 SCD 所有预期的效果,是对 DCD 有实际意义的补充.



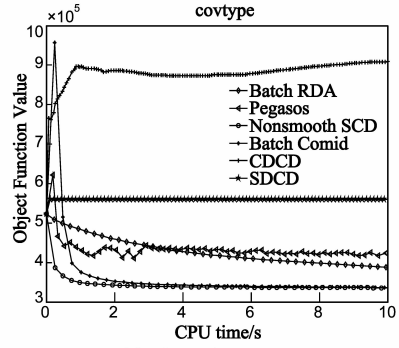
(*a*) L2-R-L1 on astro-physics



(*b*) L2-R-L1 on CCAT



(*c*) L2-R-L1 on a9a



(*d*) L2-R-L1 on covtype

图2　“L2正则化+L1损失”问题的比较实验

5　总结与展望

本文提出了一种适合求解正则化非光滑损失函数问题的 SCD 算法.理论表明这种基于 Comid 方法的 Nonsmooth SCD 算法与当前流行的算法具有同等的收敛速率,本文的实验进一步验证了算法的性能.

附录 1　式(8)解析解的求解过程

(1)如果 $P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_1$,则式(8)变为

$$w_j^{t+1}=\operatorname*{argmin}_w\{\frac{1}{\eta_t}w^2+\frac{1}{\eta_t}(g_j^t-2\eta_t w_j^t)w+\lambda|w|\} \quad (9)$$

令 $\frac{1}{\eta_t}=\frac{\beta}{2}$, $\frac{1}{\eta_t}(g_j^t-2\eta_t w_j^t)=\gamma$,式(9)变成标准的“正则化项 + 二次项 + 一次项”形式,即需要求解

$$\min_w \quad \frac{\beta}{2}w^2+\gamma w+\lambda|w| \quad (10)$$

式(10)显然是一个凸优化问题,因此 w^* 是子问题的最优解当且仅当存在 $\xi\in\partial|w^*|$ 满足 $\lambda\xi+\gamma+\beta w=0$.

$|w|$的次微分可以写成:

$$\partial|w|=\begin{cases}\{1\}, & \text{if } w>0\\ \{-1\}, & \text{if } w<0\\ \{\xi\in\mathbf{R}|-1\leqslant\xi\leqslant1\}, & \text{if } w=0\end{cases}$$

下面分三种情况讨论:

①当$|\gamma|\leqslant\lambda$ 时,一方面,如果令 $w^*=0,\xi=-\gamma/\lambda$,此时符合 $\lambda\xi+\gamma+\beta w=0$;另一方面,除此之外并无其他的解.可以通过反证得到:

如果 $w>0$,则 $\xi=1$,这时

$$\lambda\xi+\gamma+\beta w=\lambda+\gamma+\beta w>\lambda+\gamma\geqslant0$$

如果 $w<0$,则 $\xi=-1$,此时

$$\lambda\xi+\gamma+\beta w=-\lambda+\gamma+\beta w<-\lambda+\gamma\leqslant0$$

这说明对 $w\neq0$ 的两种情形,$\lambda\xi+\gamma+\beta w=0$ 均不满足.

②当 $\gamma>\lambda>0$ 时,此时 $w^*>0$ 或 $w^*=0$ 都显然不能满足(偏导 0)式的,即一定有 $w^*<0$,则 $\xi=-1$,从而 $w^*=-\frac{1}{\beta}(\gamma-\lambda)$

③当 $\gamma<-\lambda<0$ 时,类似于情形②,此时一定有 $w^*>0$,则 $\xi=1$,从而 $w^*=-\frac{1}{\beta}(\gamma+\lambda)$

以上讨论可以合并成如下形式:

$$w^*=\begin{cases}0, & \text{if}|\gamma|\leqslant\lambda\\ -\frac{1}{\beta}(\gamma-\lambda\operatorname{sgn}(\gamma)), & \text{otherwise}\end{cases}$$

将式(10)带入上式

$$w_j^{t+1}=\begin{cases}0, & \text{if}\left|\frac{g_j^t}{\eta_t}-2w_j^t\right|\leqslant\lambda\\ -\frac{\eta_t}{2}[\frac{g_j^t}{\eta_t}-2w_j^t-\lambda\operatorname{sgn}(\frac{g_j^t}{\eta_t}-2w_j^t)], & \text{otherwise}\end{cases}$$

(2)如果 $P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_2^2$,则式(8)变成

$$w_j^{t+1}=\operatorname*{argmin}_w\{(\lambda+\frac{1}{\eta_t})w^2+\frac{1}{\eta_t}(g_j^t-2\eta_t w_j^t)w\}$$

这是一个标准的二次型,解析解为,

$$w_j^{t+1}=\frac{2\eta_t w_j^t-g_j^t}{2\eta_t\lambda+2}$$

附录 2 定理 2 证明

根据 $\delta_t(\boldsymbol{W})$的定义,第 j 维是随机独立抽取,并且 $w_{j_\tau}^\tau$的值依赖于前面的选择集 $\xi_{\tau-1}$,因此有

$$\begin{aligned}\mathrm{E}_{\xi_t}\delta_t(w_{i_\tau}^\tau)&=\mathrm{E}_{\xi_t}\sum_{\tau=1}^{t}\{g_{j_\tau}^\tau(w_{j_\tau}^\tau-w_{j_\tau})+p(w_{j_\tau}^\tau)\}-\frac{1}{n}tP(\boldsymbol{W})\\&=\sum_{\tau=1}^{t}\mathrm{E}_{\xi_\tau}\{g_{j_\tau}^\tau(w_{j_\tau}^\tau-w_{j_\tau})+p(w_{j_\tau}^\tau)\}-\frac{1}{n}tP(\boldsymbol{W})\\&=\sum_{\tau=1}^{t}\mathrm{E}_{\xi_{\tau-1}}E_{j_\tau}\{g_{j_\tau}^\tau(w_{j_\tau}^\tau-w_{j_\tau})+p(w_{j_\tau}^\tau)\}-\frac{1}{n}tP(\boldsymbol{W})\end{aligned}$$

由数学期望的定义可得,

$$\begin{aligned}&\mathrm{E}_{j_\tau}\{g_{i_\tau}^\tau(w_{j_\tau}^\tau-w_{j_\tau})+p(w_{j_\tau}^\tau)\}-\frac{1}{n}P(\boldsymbol{W})\\&\quad=\frac{1}{n}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W})^{\mathrm{T}}+P(\boldsymbol{W}^\tau)]-\frac{1}{n}P(\boldsymbol{W})\\&\sum_{\tau=1}^{t}\mathrm{E}_{\xi_{\tau-1}}\mathrm{E}_{j_\tau}\{g_{j_\tau}^\tau(w_{j_\tau}^\tau-w_{j_\tau})+p(w_{j_\tau}^\tau)\}-\frac{1}{n}tP(\boldsymbol{W})\\&\quad=\frac{1}{n}\sum_{\tau=1}^{t}\mathrm{E}_{\xi_{\tau-1}}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W})^{\mathrm{T}}+P(\boldsymbol{W}^\tau)]-\frac{1}{n}tP(\boldsymbol{W})\\&\quad=\mathrm{E}_{\xi_t}\delta_t(\boldsymbol{W}^\tau)\end{aligned}$$

根据次梯度的定义,$f(\boldsymbol{W}^\tau)-f(\boldsymbol{W})\leqslant\langle\boldsymbol{g}^\tau,\boldsymbol{W}^\tau-\boldsymbol{W}\rangle$,正则化项 $P(\boldsymbol{W})$保持不变,从而

$$\begin{aligned}&\sum_{\tau=1}^{t}[f(\boldsymbol{W}^\tau)-f(\boldsymbol{W})+P(\boldsymbol{W}^\tau)]-\frac{1}{n}tP(\boldsymbol{W})\\&\quad\leqslant\sum_{\tau=1}^{t}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}+P(\boldsymbol{W}^\tau)-\frac{1}{n}tP(\boldsymbol{W})\end{aligned}$$

两边同时取期望,由 $\phi_t=\mathrm{E}_{\xi_{t-1}}[F(\boldsymbol{W}^t)]$定义得,

$$\begin{aligned}&\sum_{\tau=1}^{t}\mathrm{E}_{\xi_{\tau-1}}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}+P(\boldsymbol{W}^\tau)]-\frac{1}{n}tP(\boldsymbol{W})\\&\quad\geqslant\sum_{\tau=1}^{t}\mathrm{E}_{\xi_{\tau-1}}[f(\boldsymbol{W}^\tau)-f(\boldsymbol{W})+P(\boldsymbol{W}^\tau)]-\frac{1}{n}tP(\boldsymbol{W})\\&\quad=\frac{1}{n}\sum_{\tau=1}^{t}[\phi_\tau-F(\boldsymbol{W})]\end{aligned}$$

则有 $R_t(\boldsymbol{W})\leqslant n\mathrm{E}_{\xi_t}\delta_t(\boldsymbol{W})$

证毕.

附录 3 定理 3 的证明

(1)$P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_1$,令 $\boldsymbol{W}=\boldsymbol{W}^*$,$\boldsymbol{r}\in\partial P(\boldsymbol{W})$,根据次梯度的定义则有,

$$\begin{aligned}&n\mathrm{E}_{\xi_t}\delta_t(\boldsymbol{W}^\tau)\\&\quad=\sum_{\tau=1}^{t}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}+P(\boldsymbol{W}^\tau)-P(\boldsymbol{W}^*)]\\&\quad\leqslant\sum_{\tau=1}^{t}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}+r^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}]\\&\sum_{\tau=1}^{t}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{T}+\boldsymbol{r}^{\tau+1}(\boldsymbol{W}^{\tau+1}-\boldsymbol{W}^*)^{\mathrm{T}}]\\&\quad-\boldsymbol{r}^{t+1}(\boldsymbol{W}^{t+1}-\boldsymbol{W}^*)^{\mathrm{T}}+r^1(\boldsymbol{W}^1-\boldsymbol{W}^*)^{\mathrm{T}}\\&\quad=\sum_{\tau=1}^{t}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}+\boldsymbol{r}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}]\end{aligned}$$

根据假设 $\|\boldsymbol{r}\|\leqslant M$,$B_\varphi(\boldsymbol{W},\boldsymbol{W}^t)\leqslant D^2$ 且 $\frac{\alpha}{2}\|\boldsymbol{W}-\boldsymbol{W}^t\|^2\leqslant B_\varphi(\boldsymbol{W},\boldsymbol{W}^t)$可得

$$\begin{aligned}-\sqrt{\frac{2}{\alpha}}DM&\leqslant\boldsymbol{r}^1(\boldsymbol{W}^1-\boldsymbol{W}^*)^{\mathrm{T}}\\&\leqslant\|\boldsymbol{r}^1\|\,\|\boldsymbol{W}^1-\boldsymbol{W}^*\|\leqslant\sqrt{\frac{2}{\alpha}}DM,\\-\sqrt{\frac{2}{\alpha}}DM&\leqslant\boldsymbol{r}^{t+1}(\boldsymbol{W}^{t+1}-\boldsymbol{W}^*)^{\mathrm{T}}\\&\leqslant\|\boldsymbol{r}^{t+1}\|\,\|\boldsymbol{W}^{t+1}-\boldsymbol{W}^*\|\leqslant\sqrt{\frac{2}{\alpha}}DM\end{aligned}$$

当 $P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_1$,令 $\eta_\tau=\frac{D\sqrt{\alpha}}{G_*\sqrt{\tau}}$,根据文献[17]一般凸条件下收敛性证明,有下式:

$$\begin{aligned}&\sum_{\tau=1}^{t}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}+\boldsymbol{r}^\tau(\boldsymbol{W}^{\tau+1}-\boldsymbol{W}^*)^{\mathrm{T}}]\leqslant\frac{DG_*^2\sqrt{t}}{\sqrt{\alpha}}+2\sqrt{\frac{2}{a}}DG_*\\&\sum_{\tau=1}^{t}[\boldsymbol{g}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}+\boldsymbol{r}^\tau(\boldsymbol{W}^\tau-\boldsymbol{W}^*)^{\mathrm{T}}]\\&\quad\leqslant\frac{DG_*^2\sqrt{t}}{\sqrt{\alpha}}-\boldsymbol{r}^{t+1}(\boldsymbol{W}^{t+1}-\boldsymbol{W}^*)^{T}+\boldsymbol{r}^1(\boldsymbol{W}^1-\boldsymbol{W}^*)^{\mathrm{T}}\\&\quad\leqslant\frac{DG_*^2\sqrt{t}}{\sqrt{\alpha}}+2\sqrt{\frac{2}{a}}DG_*+2\sqrt{\frac{2}{\alpha}}DM=\frac{DG_*^2\sqrt{t}}{\sqrt{\alpha}}+\mathrm{constant}\end{aligned}$$

由附录 1 的结论可得,

$$\sum_{\tau=1}^{t}[\varphi_\tau-F(\boldsymbol{W}^*)]\leqslant n\mathrm{E}_{\xi_t}\delta_t(\boldsymbol{W})\leqslant\frac{DG_*^2\sqrt{t}}{\sqrt{\alpha}}+\mathrm{constant}$$

则 $\frac{1}{t}\sum_{\tau=1}^{t}\phi_\tau-F(\boldsymbol{W}^*)\leqslant O(\frac{DG_*^2\sqrt{t}}{\sqrt{\alpha}t})$

(2)若 $P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_2^2$,令 $\eta_\tau=\frac{1}{\lambda\tau}$,根据文献[17]强凸条件下收敛性证明,同理可得,

$$\frac{1}{t}\sum_{\tau=1}^{t}\phi_\tau-F(\boldsymbol{W}^*)\leqslant O(\frac{G_*^2}{\lambda a}\frac{\ln t}{t})$$

证毕.

参考文献

[1] 孙正雅,陶卿.统计机器学习综述:损失函数与优化求解[J].中国计算机学会通讯,2009,5(8):7-14.
Sun Zheng-Ya, Tao Qing. Statistical machine learning: A review of the loss function and optimization [J]. Communications of the CCF, 2009, 5(8): 7-14. (in Chinese)

[2] Y Nesterov. Effciency of coordinate descent methods on huge-scale optimizationproblems [R]. Core discussion papers, (0022010), 2010.

[3] K W Chang, C J Hsieh, C J Lin, Coordinate descent method for large-scale L2-loss linear support vector machines[J]. Journal of Machine Learning Research, 2008, 9: 1369-1398.

[4] C J Hsieh, K W Chang, C J Lin, S S Keerthi, S Sundararajan. A

dual coordinate descent method for large-scale linear SVM [A]. Proceedings of the 25th International Conference on Machine Learning[C]. USA: ACM, 2008. 408 – 415.

[5] S P Boyd, L Vandenberghe. Convex optimization[M]. England: Cambridge University Press, 2004. 121 – 139.

[6] P Tseng. Convergence of a block coordinate descent method for non-differentiable minimization[J]. Journal of optimization theory and applications, 2001, 109(3): 475 – 494.

[7] P Tseng, S Yun. A coordinate gradient descent method for nonsmooth separable minimization [J]. Mathematical Programming, 2009, 117(1): 387 – 423.

[8] 焦李成,杨淑媛,等.压缩感知回顾与展望[J].电子学报, 2011, 39(7): 1651 – 1662.
Jiao Li-Cheng, Yang Shu-Yuan, et al. Development and Prospect of Compressive Sensing[J]. Acta Electronica Sinica, 2011, 39(7): 1651 – 1662. (in Chinese)

[9] S Shalev-Shwartz, A Tewari. Stochastic methods for l1 regularized loss minimization[A]. Proceedings of the 26th International Conference on Machine Learning[C]. USA: ACM, 2009. 929 – 936.

[10] S Yun, K C Toh. A coordinate gradient descent method for L1-regularized convex minimization[J]. Computational Optimization and Applications, 2011, 48(2): 273 – 307.

[11] A Saha, A Tewari. On the fnite time convergence of cyclic coordinate descent methods[R]. Technical report, 2010.

[12] A Beck, M Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems[J]. SIAM Journal on Imaging Sciences, 2009, 2(1): 183 – 202.

[13] G X Yuan, K W Chang, C J Hsieh, C Lin. A comparison of optimization methods and software for large-scale L1-regularized linear classifcation[J]. Journal of Machine Learning Research, 2010, 11: 3183 – 3234.

[14] Y Nesterov. How to advance in structural convex optimization [J]. OPTIMA, MPS Newsletter, 2008, 78: 2 – 5.

[15] Y Nesterov. Primal-dual subgradient methods for convex problems[J]. Mathematical programming, 2009, 120(1): 221 – 259.

[16] L Xiao. Dual averaging methods for regularized stochastic learning and online optimization [J]. Journal of Machine Learning Research, 2010, 11: 2543 – 2596.

[17] J Duchi, S Shalev-Shwartz, Y Singer, A Tewari. Composite objective mirror descent[A]. Proceedings of the 23th Annual Workshop on Computational Learning Theory [C]. USA: ACM, 2010. 116 – 128.

[18] J Duchi, S Shalev-Shwartz, Y Singer, T Chandra. Efficient projections onto the L1-ball for learning in high dimensions [A]. Proceedings of the 25th International Conference on Machine Learning[C]. USA: ACM, 2008. 272 – 279.

[19] Tao Qing, Sun Zheng-ya, Kong Kang. developing learning algorithms via optimized discretization of continuous dynamical systems[J]. IEEE Trans Syst Man Cybern B, 2012, 42(1): 140 – 149.

[20] A Beck, M Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization[J]. Operations Research Letters, 2003, 31(3): 167 – 175.

[21] E Hazan, A Kalai, S Kale, A Agarwal. Logarithmic regret algorithms for online convex optimization[A]. Proceedings of the 19th Annual Workshop on Computational Learning Theory [C]. USA: ACM, 2006. 499 – 513.

[22] M Zinkevich. Online convex programming and generalized infinitesimal gradient ascent[A]. Proceedings of the 20th International Conference on Machine Learning[C]. USA: ACM, 2003. 159 – 166.

[23] S Shalev-Shwartz, Y Singer, N Srebro. Pegasos: Primal estimated sub-gradient solver for SVM[A]. Proceedings of the 24th International Conference on Machine Learning [C]. USA: ACM, 2007. 807 – 814.

[24] T Joachims. Training linear SVMs in linear time[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C]. USA: ACM, 2006. 217 – 226.

[25] A J Smola, S V N Vishwanathan, Q Le. Bundle methods for machine learning[A]. Advances in Neural Information Processing Systems[C]. USA: MIT Press, 2008, 20: 1377 – 1384.

[26] C J Lin, R C Weng, S S Keerthi. Trust region Newton method for large-scale logistic regression [J]. Journal of Machine Learning Research, 2008, 9: 627 – 650.

作者简介

陶　卿　男,1965 年生于安徽.博士,教授,中国计算机学会高级会员,中国科学院自动化研究所博士生导师,主要研究方向为统计机器学习与人工智能.
E-mail: qing.tao@ia.ac.cn

朱烨雷　男,1985 年生于江苏.硕士研究生,主要研究方向为模式识别与人工智能、数据挖掘、图像处理.