

一种相关快速软阈值坐标下降算法

王玉军

(中国人民解放军陆军军官学院,安徽 合肥 230031)

摘要:软阈值缩减迭代算法(ISTA)以其简单的操作流程成为了机器学习流行的优化算法,但是收敛速度比较慢,仅为 $o(\frac{1}{k})$ 。快速软阈值缩减迭代算法(FISTA)通过加速技巧将收敛速度提高了一个数量级,达到了 $o(\frac{1}{k^2})$ 。然而,FISTA 将特征向量每一维看成是独立同分布的,丢失了各维之间的相关性,会导致准确率下降和额外的时间开销。为了弥补上述的不足,文中提出了一种相关快速软阈值坐标下降算法(RFTCD)。通过大规模数据库实验证实了 RFTCD 的正确性和有效性。

关键词:软阈值缩减迭代;机器学习;特征向量;独立同分布;坐标下降

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2013)12-0055-04

doi:10.3969/j.issn.1673-629X.2013.12.013

A Relative Fast Soft-thresholding Coordinate Descent Algorithm

WANG Yu-jun

(Chinese People's Liberation Army Officer Academy, Hefei 230031, China)

Abstract: Although iterative shrinkage-thresholding algorithm (ISTA) becomes popular optimization algorithms of machine learning because of its simple operational processes, but the convergence rate is slow, only $o(\frac{1}{k})$. Convergence rate of fast iterative shrinkage-thresholding algorithm (FISTA) by accelerating skills can improve by an order of magnitude, reaching $o(\frac{1}{k^2})$. However each eigenvectors dimension is seen by FISTA as independent and identically distributed, which will loss the correlation between each dimension and lead to the decline in accuracy and time overhead. In order to circumvent these drawbacks, present a relative fast soft-thresholding coordinate descent algorithm. Extensive experiments on large-scale real database verify the proposed algorithm is correct and effective.

Key words: iterative shrinkage-thresholding; machine learning; feature vector; independent and identically distributed; coordinate descent

0 引言

大多数的统计机器学习问题可以表示成为“正则化项+损失函数”形式的优化问题^[1-4]。具体来说,给定训练样本集 $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \in R^n \times \{-1, +1\}$, 一个二分类问题的求解可描述为求解下述正则化形式的优化问题:

$$\min F(\mathbf{w}) = \lambda f(\mathbf{w}) + g(\mathbf{w}) \quad (1)$$

其中, $f(\mathbf{w})$ 为正则化项; $g(\mathbf{w})$ 称为损失函数;参数 λ 则反映了这两者之间的折中。文中主要研究 $f(\mathbf{w}) = \sum_{i=1}^n |w_i|$, $g(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m (\mathbf{w}^T \mathbf{x}_i - y_i)^2$ 是光滑的。则原优化问题的具体形式如下:

$$\min F(\mathbf{w}) = \lambda \sum_{i=1}^n |w_i| + \frac{1}{m} \sum_{i=1}^m (\mathbf{w}^T \mathbf{x}_i - y_i)^2 \quad (2)$$

问题2就是熟悉的Lasso问题,它是强凸光滑的。当前求解问题2有多种方法,文献[5]利用其二阶梯度信息使用内点法求解,但是在许多实际应用中,比如图像去模糊化,需要处理大规模的数据,而且数据是稠密的,内点法需要巨大的时间开销,效率低下。在处理大规模数据的方法中,比较流行的是文献[6-8]提出的软阈值缩减迭代算法(Iterative Shrinkage-Thresholding Algorithm, ISTA),该算法以其操作过程简单而很受欢迎,收敛速度为 $o(\frac{1}{k})$,但是已经被^[9-10]认为是一种慢的算法,尤其是在一些特定的假设下^[9], ISTA

收稿日期:2013-03-04

修回日期:2013-06-09

网络出版时间:2013-09-29

基金项目:国家自然科学基金资助项目(60975040)

作者简介:王玉军(1984-),男,江苏盐城人,硕士研究生,研究方向为模式识别与人工智能、数据挖掘、图像处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130929.1541.032.html>

在处理图像去模糊化中算法收敛是比较慢的,去模糊的效果也不是很理想。为了解决 ISTA 的问题,2009 年 Amir Beck^[11] 提出了快速软阈值缩减迭代算法 (Fast Iterative Shrinkage-Thresholding Algorithm, FISTA), FISTA 既保持了 ISTA 的简单性,同时通过加速的方法将收敛速度提高到了 $O\left(\frac{1}{k^2}\right)$, 在时间开销上取得了数量级的提升,然而 FISTA 将特征向量每一维看成独立同分布的,对各维单独处理,丢失了各维之间的相关性,导致准确率下降和额外的时间开销。文中就是在这一背景下,提出了一种相关快速软阈值坐标下降 (Relative Fast Thresholding Coordinate Descent, RFTCD) 算法。

目前,坐标下降方法以低廉的计算代价、快速的实际收敛效果和简洁的操作流程,吸引了众多研究者的关注。针对比较常用的 L1 正则化最小二乘回归 (Lasso) 问题,1998 年, Fu W J^[12] 提出了用坐标下降加以求解的办法。到目前为止,坐标下降方法的强大优势才得以充分的发掘。文献[13]指出如果能够合理地实现坐标下降方法,将可以得到比当前主流算法快的多的算法。

2008 年,林智仁教授的研究小组对多种不同损失函数的学习问题建立了原始和对偶坐标下降方法^[14],实验结果表明坐标优化方法能够充分利用高维数据的稀疏特性,取得了比信赖域方法^[15]、Pegasos^[16]、割平面方法^[17]更好的效果,坐标优化方法已经成为处理大规模稀疏数据特别是文本数据的首选算法。

文中就是在坐标优化的基础上提出了 RFTCD 算法,在大规模数据库上取得了预期的效果。

1 基本理论

1.1 坐标下降 (Coordinate Descent, CD) 方法

原始问题的坐标下降方法是坐标下降方法中一种最常见形式,它将 \mathbf{w} 的每一维看成是一个坐标,其迭代的过程分为内循环和外循环两部分^[18]。迭代过程从起始点 \mathbf{w}^0 开始依次迭代出 $\mathbf{w}^1, \mathbf{w}^2, \dots, \mathbf{w}^T$ 。从 \mathbf{w}^t 到 \mathbf{w}^{t+1} 的过程称为一次外循环。 \mathbf{w}^{t+1} 通过更新 \mathbf{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 $\mathbf{w}^{t,j} \in R^n, j = 1, \dots, n$, 并且 $\mathbf{w}^{t,1} = \mathbf{w}^t, \mathbf{w}^{t,n+1} = \mathbf{w}^{t+1}, \mathbf{w}^{t,j} = [w_1^{t+1}, \dots, w_{j-1}^{t+1}, w_j^t, \dots, w_n^t]^T, j = 2, \dots, n$, 首尾相接。对于 $\mathbf{w}^{t,j}$ 到 $\mathbf{w}^{t,j+1}$ 的更新,通过求解如下单变量子问题得到:

$$\min_z \mathbf{w}^{t,j} = [w_1^{t+1}, \dots, w_{j-1}^{t+1}, w_j^t + z, w_{j+1}^t, \dots, w_n^t]^T = \min_z F(\mathbf{w}^{t,j} + z\mathbf{e}_j), \mathbf{e}_j = [0, \dots, 1, \dots, 0]^T \quad (3)$$

坐标下降方法的流程见算法 1。

算法 1: 坐标下降方法。

Start with any initial \mathbf{w}^0

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

For $j = 1, 2, \dots, n$ (inter iterations)

Fix $w_1^{t+1}, \dots, w_{j-1}^{t+1}, w_{j+1}^t, \dots, w_n^t$ and

approximately solve the sub-problem(3)

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子问题的求解方式上。对于问题 2 的子问题为

$$\min_z g(\mathbf{w}^{t,j} + z\mathbf{e}_j) + \lambda \|\mathbf{w}^{t,j} + z\mathbf{e}_j\|_1 \quad (4)$$

1.2 Lipschitz 理论

定义 1^[19-20]: 如果存在常数 $L > 0$, 使得对定义域 D 的任意两个不同的实数 W, V 均有:

$$\|\nabla\theta(W) - \nabla\theta(V)\| \leq L \|W - V\| \quad (5)$$

其中, L 是 Lipschitz 常数。在机器学习中一般将满足上述条件的损失称之为光滑损失函数,对于不满足的则是非光滑损失函数。问题 2 中 $f(\mathbf{w})$ 是非光滑函数, $g(\mathbf{w})$ 是光滑函数。

引理 1: 如果 $\theta(\mathbf{x})$ 是连续可微光滑函数, Lipschitz 常数为 $L(\theta)$, 对于任意 $L \geq L(\theta), \mathbf{x}, \mathbf{y} \in R^n$, 则有:

$$\theta(\mathbf{x}) \leq \theta(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \nabla\theta(\mathbf{y}) \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (6)$$

将式(2)中 $g(\mathbf{w}^{t+1})$ 在 \mathbf{w}^t 处 Lipschitz 近似展开:

$$Q_L(\mathbf{w}^{t+1}, \mathbf{w}^t) = \lambda \sum_{j=1}^n |w_j^{t+1}| + \langle \mathbf{w}^{t+1} - \mathbf{w}^t, \nabla g(\mathbf{w}^t) \rangle + \frac{L}{2} \|\mathbf{w}^{t+1} - \mathbf{w}^t\|^2 \quad (7)$$

则原优化问题转化为如下优化问题:

$$p_L(\mathbf{w}^t) = \operatorname{argmin}_{\mathbf{w}} \{Q_L(\mathbf{w}^{t+1}, \mathbf{w}^t), \mathbf{w}^{t+1} \in R^n\} \quad (8)$$

式(7)中去除关于 \mathbf{w}^t 的常数项,

$$p_L(\mathbf{w}^t) = \operatorname{argmin}_{\mathbf{w}} \left\{ \lambda \sum_{j=1}^n |w_j^{t+1}| + \frac{L}{2} \|\mathbf{w}^{t+1} - (\mathbf{w}^t - \frac{1}{L} \nabla g(\mathbf{w}^t))\|^2 \right\} \quad (9)$$

2 ISTA 和 FISTA

ISTA 和 FISTA 都是坐标优化算法,其单变量子问题为:

$$\lambda |w_j^t + z| + (\nabla g_j(\mathbf{w}^t) - Lw_j^t)(w_j^t + z) + \frac{L}{2} (w_j^t + z)^2 \quad (10)$$

其中, $\nabla g_j(\mathbf{w}^t) = \sum_{i=1}^m (\langle \mathbf{w}^t, \mathbf{x}_i \rangle - y_i) \cdot \mathbf{x}_i^j$ 。

式(10)满足“正则化项 + 一次项 + 二次项”结构,可以使用软阈值方法求解^[15],最小值即为当前子问题的解析解(closed-form solution)。

$$z^* = \begin{cases} -w_j^t, & \text{if } |\alpha| \leq \lambda \\ -\frac{\nabla g_j(\mathbf{w}^t) - \lambda \operatorname{sgn}(\theta_j)}{L}, & \text{otherwise} \end{cases}$$

其中, $\alpha = \nabla g_j(\mathbf{w}^t) - Lw_j^t$ 。

ISTA 具体的流程见算法 2。

算法 2:ISTA。

Tuning parameters: λ, L ;

Initialize a weight vector $\mathbf{w}_0 = \mathbf{0}$;

For $t = 1, 2, \dots$

compute $\nabla g(\mathbf{w}_t)$;

For $j = 0, 1, \dots, n$

1) compute $\alpha = \nabla g_j(\mathbf{w}^t) - Lw_j^t$;

2) z^* 由式(10)求得;

3) $w_j^{t+1} = w_j^t + z^*$;

End

从算法 2 可以看出, ISTA 流程简单、操作方便, 但是收敛速度比较慢, 仅为 $o\left(\frac{1}{k}\right)$ 。为了提高收敛速度, 文献[8]提出了 FISTA, 具体的流程见算法 3。

算法 3:FISTA。

Tuning parameters: λ, L ;

Initialize a weight vector $\mathbf{y}_0 = \mathbf{w}_0 = \mathbf{0}, k = 1$;

For $t = 1, 2, \dots$

compute $\nabla g(\mathbf{w}_t)$;

For $j = 0, 1, \dots, n$

1) compute $\alpha = \nabla g_j(\mathbf{w}^t) - Lw_j^t$;

2) z^* 由式(10)求得;

3) $w_j^{t+1} = w_j^t + z^*$;

4) $k_{t+1} = \frac{1 + \sqrt{1 + 4k_t}}{2}$;

5) $y_j^{t+1} = w_j^{t+1} + \left(\frac{k_t - 1}{k_{t+1}}\right)(w_j^{t+1} - w_j^t)$;

End

算法 3 和算法 2 不同之处在于算法 2 内循环中每维的更新即步骤 3 只用到了上一个阶段 w_j^t , 而算法 3 通过引入中间变量 \mathbf{y} , 每维更新不仅用到了上一个阶段 w_j^t , 还用到了当前阶段 w_j^{t+1} , 更新时用的信息充实而合理, 这就是 FISTA 算法收敛性优于 ISTA 算法的原因, 达到了 $o\left(\frac{1}{k^2}\right)$ 。两个算法的共同之处: 每维的更新没有用到上一维的信息, 主要表现在计算 $\nabla g(\mathbf{w}_{t-1})$, 是在内循环更新每维前一起更新的。这将特征向量每一维看成是独立同分布的, 丢失了各维之间的相关性, 会导致准确率下降和额外的时间开销。

3 RFTCD 算法

为了弥补上述两个算法各维更新缺失的相关性, 文中提出了一种相关快速软阈值坐标下降算法。具体

的算法流程如下。

算法 4:RFTCD 算法。

Tuning parameters: λ, L ;

Initialize a weight vector $\mathbf{y}_0 = \mathbf{w}_0 = \mathbf{0}, k = 1$;

For $t = 1, 2, \dots$

For $j = 0, 1, \dots, n$

1) compute $\nabla g_j(\mathbf{w}_t)$;

2) compute $\alpha = \nabla g_j(\mathbf{w}^t) - Lw_j^t$;

3) z^* 由式(10)求得;

4) $w_j^{t+1} = w_j^t + z^*$;

5) $k_{t+1} = \frac{1 + \sqrt{1 + 4k_t}}{2}$;

6) $y_j^{t+1} = w_j^{t+1} + \left(\frac{k_t - 1}{k_{t+1}}\right)(w_j^{t+1} - w_j^t)$;

End

RFTCD 算法和 FISTA 不同之处在于步骤 1, 每维梯度是在内循环中更新的, 方法很简单但可以使每一维的更新是建立在上一维的基础上的, 充分地利用了各维的相关性, 可以有效地提高准确率和减少时间开销。

4 实验结果

实验在 Sun Ultra45 工作站 (1.6 GHz UltraSPARC IIIi 处理器, 4 GB 内存, Solaris10 操作系统) 上实现。该实验采用标准 C/C++ 语言实现。

文中用到 3 个大规模数据库 astro-physic, ccat, real-sim, 它们都属于文本分类库 (见表 1)。3 个数据库的下载地址为 <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>。

表 1 3 个大规模库描述

Dataset	Training Samples	Test Samples	Dimension
astro-physic	29,882	32,487	99,757
ccat	23,149	781,265	47,236
real-sim	30,000	42,309	20,958

实验结果见表 2, 其中, accuracy: 测试精度, time: 算法的训练时间 (毫秒)。

表 2 大规模数据库实验比较

Dataset	ISTA		FISTA		RFTCD	
	accuracy	time/	accuracy	time/	accuracy	time
	/%	ms	/%	ms	/%	/ms
astro-physic	95.17	4 506	95.20	3 910	96.31	3 802
ccat	92.51	2 086	92.49	1 736	93.18	1 658
real-sim	96.89	2 316	96.95	1 932	97.77	1 892

5 结束语

文中在坐标优化的基础上提出了一种相关快速软阈值坐标下降算法 (RFTCD)。通过大规模数据库实

验证了 RFTCD 的正确性和有效性。

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.
- [2] 周志华,杨强.机器学习及其应用 2011[M].北京:清华大学出版社,2011.
- [3] 边肇祺,张学工.模式识别[M].北京:清华大学出版社,2000.
- [4] 周志华.机器学习[J].中国计算机学会通讯,2009,5(8):6-6.
- [5] Ben-Tal A,Nemirovski A. Lectures on modern convex optimization: Analysis, algorithms, and engineering applications [M]. Philadelphia:SIAM,2001.
- [6] Chambolle A,de Vore R A, Lee N Y, et al. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage[J]. IEEE trans on image processing, 1998,7(3):319-335.
- [7] Figueiredo M A T, Nowak R D. An EM algorithm for wavelet-based image restoration[J]. IEEE trans on image processing, 2003,12(8):906-916.
- [8] Daubechies I, Defrise M, Mol C D. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint [J]. Comm pure appl math, 2004,57(11):1413-1457.
- [9] Bredies K, Lorenz D. Iterative soft-thresholding converges linearly[R/OL]. 2008. <http://arxiv.org/abs/0709.1598v3>.
- [10] Bruck R J. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space[J]. J math anal appl, 1977,61(1):159-164.
- [11] Beck A, Teboulle M. A fast iterative shrinkage-thresholding

algorithm for linear inverse problems[J]. Society for industrial and applied mathematics, 2009,2(3):183-202.

- [12] Fu W J. Penalized regressions: The bridge versus the lasso [J]. Journal of computational and graphical statistics, 1998,7(3):397-416.
- [13] Friedman J, Hastie T, Tibshirani R, et al. Pathwise coordinate optimization[J]. The annals of applied statistics, 2007,1(2):302-332.
- [14] Chang Kaiwei, Hsieh C J, Lin C J. Coordinate descent method for large-scale L2-loss linear support vector machines[J]. Journal of machine learning research, 2008,49(7):1369-1398.
- [15] Lin C J, Ruby C W, Keerthi S. Trust region Newton method for large-scale logistic regression[J]. Journal of machine learning research, 2008,9(10):627-650.
- [16] Shalev-Shwartz S, Singer Y, Srebro N. Pegasos: Primal estimated sub-gradient solver for SVM[C]//Proc of the 24th international conference on machine learning. New York:ACM, 2007:807-814.
- [17] Franc V, Sonnenburg S. Optimized cutting plane algorithm for support vector machines[C]//Proc of the 25th international conference on machine learning. New York:ACM, 2008:320-327.
- [18] Xiao Ling. Dual averaging methods for regularized stochastic learning and online optimization[J]. The journal of machine learning research, 2010,3(11):2543-2569.
- [19] 陈宝林.最优化理论与算法[M].第2版.北京:清华大学出版社,2005.
- [20] 袁亚湘,孙文瑜.最优化理论与方法[M].北京:科学出版社,2007.

(上接第 54 页)

- [2] 金玲玲,汪文俊,王喜凤.大学生综合素质的灰色模糊聚类评价模型[J].计算机技术与发展,2012,22(5):109-112.
- [3] 任永昌,彭霞,常革新.软件项目质量控制相关技术研究[J].计算机技术与发展,2012,22(10):143-146.
- [4] 任永昌.软件项目管理[M].北京:清华大学出版社,2012.
- [5] 百度百科.软件性能[EB/OL]. 2012-12-01. <http://baike.baidu.com/view/1812806.html>.
- [6] 百度知道.简述信息安全的重要性[EB/OL]. 2012-12-01. <http://zhidao.baidu.com/question/337046357.html>.
- [7] Na K S, Simpson J T, Li Xiaotong, et al. Software development risk and project performance measurement: Evidence in Korea [J]. Journal of systems and software, 2007,80(4):596-605.
- [8] 邹珊刚,唐炎钊.投资项目的灰色综合评价及应用[J].华中理工大学学报,1999,27(7):92-94.
- [9] MBA 智库百科.权重的设定方法[EB/OL]. 2012-12-01. <http://wiki.mbalib.com/wiki/权重>.
- [10] 吴祈宗.系统工程[M].北京:北京理工大学出版社,2006.
- [11] Baskaran V, Nachiappan S, Rahman S. Indian textile suppliers

' sustainability evaluation using the gray approach [J]. International journal of production economics, 2012,135(2):647-658.

- [12] Zhou J G, Wang Y X, Li B. Study on optimization of denitration technology based on gray-fuzzy combined comprehensive evaluation model [J]. Systems engineering procedia, 2012,4(1):210-218.
- [13] 张晓明.决策分析中的数据无量纲化方法比较分析[J].闽江学院学报,2012,33(5):21-25.
- [14] 叶宗裕.关于多指标综合评价中指标正向化和无量纲化方法的选择[J].浙江统计,2003,22(4):24-25.
- [15] 张晓丰,郭建胜,张凤鸣.软件方案选择灰局势决策[J].微电子学与计算机,2006,23(1):105-107.
- [16] Ren Y C, Xing T, Liu D C. Establishment of comprehensive capacity evaluation index system on system analyst[C]//Proc of 10th conference on man-machine-engineering. USA:Scientific Research Publishing, 2010:43-47.

一种相关快速软阈值坐标下降算法

作者：[王玉军](#)，[WANG Yu-jun](#)
作者单位：[中国人民解放军陆军军官学院, 安徽 合肥, 230031](#)
刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

ISTIC

年, 卷(期):[2013\(12\)](#)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjfz201312013.aspx