软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2498−2507 [doi: 10.3724/SP.J.1001.2013.04479]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

稀疏学习优化问题的求解综述*

陶 卿, 高乾坤, 姜纪远, 储德军

(中国人民解放军陆军军官学院 11 系,安徽 合肥 230031)

通讯作者: 陶卿, E-mail: taoqing@gmail.com

摘 要: 机器学习正面临着数据规模日益扩大的严峻挑战,如何处理大规模甚至超大规模数据问题,是当前统计学习亟需解决的关键性科学问题.大规模机器学习问题的训练样本集合往往具有冗余和稀疏的特点,机器学习优化问题中的正则化项和损失函数也蕴含着特殊的结构含义,直接使用整个目标函数梯度的批处理黑箱方法不仅难以处理大规模问题,而且无法满足机器学习对结构的要求.目前,依靠机器学习自身特点驱动而迅速发展起来的坐标优化、在线和随机优化方法成为解决大规模问题的有效手段.针对 L1 正则化问题,介绍了这些大规模算法的一些研究进展.

关键词: L1 正则化;在线优化;随机优化;坐标优化

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 陶卿,高乾坤,姜纪远,储德军.稀疏学习优化问题的求解综述.软件学报,2013,24(11):2498−2507. http://www.jos.org.cn/1000-9825/4479.htm

英文引用格式: Tao Q, Gao QK, Jiang JY, Chu DJ. Survey of solving the optimization problems for sparse learning. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2498−2507 (in Chinese). http://www.jos.org.cn/1000-9825/4479.htm

Survey of Solving the Optimization Problems for Sparse Learning

TAO Qing, GAO Qian-Kun, JIANG Ji-Yuan, CHU De-Jun

(11th Department, Army Officer Academy of PLA, Hefei 230031, China)

Corresponding author: TAO Qing, E-mail: taoqing@gmail.com

Abstract: Machine learning is facing a great challenge arising from the increasing scale of data. How to cope with the large-scale even huge-scale data is a key problem in the emerging area of statistical learning. Usually, there exist redundancy and sparsity in the training set of large-scale learning problems, and there are structural implications in the regularizer and loss function of a learning problem. If the gradient-type black-box methods are employed directly in batch settings, not only the large-scale problems cannot be solved but also the structural information implied by the machine learning cannot be exploited. Recently, the state-of-the-art scalable methods such as coordinate descent, online and stochastic algorithms, which are driven by the characteristics of machine learning, have become the dominant paradigms for large-scale problems. This paper focuses on L1-regularized problems and reviews some significant advances of these scalable algorithms.

Key words: L1-regularization; online optimization; stochastic optimization; coordinate optimization

自 1995 年以来,统计机器学习的理论分析经历了“间隔”和“损失函数”两个重要的发展阶段,在理论分析工具方面已经基本成熟[1−4].众所周知,在损失函数具有贝叶斯一致性的前提条件下,目前大多数的机器学习算法都遵循正则化经验损失的设计框架.在具体应用中,一旦一个学习问题的损失函数确定以后,所面临的任务主要是如何求解正则化损失函数导致的优化问题[4].目前,机器学习和优化理论的研究者已经将多种不同优化原理的算法引入到机器学习领域,取得了显著的效果[5].随着计算机和网络技术的飞速发展,机器学习所面临的数据

* 基金项目: 国家自然科学基金(60975040, 61273296); 安徽省自然科学基金(1308085QF121)
收稿时间: 2013-04-30; 修改时间: 2013-07-16, 2013-08-02; 定稿时间: 2013-08-27

规模正变得越来越大,一个普通文本数据库就会达到 10^{10} 样本个数或 10^{10} 样本维数的规模.显然,如何有效求解这种大规模正则化损失函数优化问题,是目前机器学习亟需解决的科学问题.

1 L1 正则化学习问题

本文仅考虑最简单的一种典型稀疏学习问题,即下面具有 L1 正则化项的二分类问题:

$$\lambda \| \boldsymbol{w} \|_1 + \frac{1}{m}\sum_{i=1}^{m} l(\boldsymbol{w},(\boldsymbol{x}_i,y_i)) \tag{1}$$

其中,训练样本集 $S=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_m,y_m),(\boldsymbol{x}_i,y_i)\in R^n\times\{-1,+1\}\}$ 由独立同分布的样本组成,$l(\boldsymbol{w}(\boldsymbol{x}_i,y_i))$ 为样本 $(\boldsymbol{x}_i,y_i)$ 的损失.从稀疏性定义的角度来说,正则化项为 L_0 范数的优化问题(1)才是原始的稀疏学习问题,但此时优化问题是 NP 完全的,目前仍无有效方法求解.从逼近和方便计算的角度来说,人们采用 L1 正则化项代替了导致问题的 L_0 正则化项.这主要是因为:一方面,L1 正则化导致的优化问题是凸的,从而可以有效求解;另一方面,如果实际问题的解具有稀疏性,那么采用 L1 正则化项的方法在一定条件下确实能够得到这种稀疏解[6].

目前,机器学习领域所使用的术语“稀疏学习问题”实际上指的是使用逼近正则化项的优化问题.

由于公式(1)是一个无约束的凸优化问题,人们首先想到的是直接使用来源于凸分析的批处理优化算法进行求解,如使用梯度下降方法和内点法就能得到很好的效果,但如果将机器学习中正则化损失函数优化问题完全归结于一个数学规划问题来研究,特别是对大规模问题,无论是求解方法还是解的含义,往往会无法满足机器学习的实际需求.这主要是因为:

① 由于求解正则化损失函数的梯度需要知道每一个样本的信息,因此批处理梯度方法的每一步迭代都不得不遍历所有的样本,同时又要使用所有维的信息,这种求解方式构成了处理大规模数据的主要瓶颈.

② 数学规划问题的数据来源往往十分确切,优化理论的方法重点关注算法的收敛速度与精度,而在统计学习问题中,实际训练数据存在着不确定性和噪声.由于统计学习算法的解与贝叶斯解之间不可避免地存在着一些差异,因此统计学习的优化算法具有完全不同的评价标准,即无须具有很快的理论收敛速度或者以很高的精度收敛,只要能够很快获得泛化性能较好的模型且计算开销小即可.

③ 大规模机器学习问题的数据来源具有一定的特点:首先,样本是独立同分布,在一些情形下,样本集合存在着冗余现象,少部分样本甚至极少一部分样本已足以反映样本集合的统计规律;其次,样本是稀疏的,部分甚至绝大部分属性是 0.如果不充分利用这些特点,完全依赖黑箱批处理方法直接求解,往往会由于硬件条件的限制而使优化算法无法运行,或者由于求解时间的局限无法满足实际需求.

为了解决黑箱批处理方法的存储和计算代价问题,机器学习领域目前主要采用一次处理一个样本点或者一次处理一维坐标的方法,这就是随机和坐标优化方法的主要思想.这些方法能够充分利用学习问题的特点,已经迅速发展成为解决大规模问题的有效手段.值得指出的是,这些方法往往是已有批处理优化方法的拓广,但机器学习优化问题中的正则化项和损失函数往往有着特殊的含义.对于这种具备学习特点的优化问题,如果直接使用已有的黑箱优化方法不加区别地作用于整个目标函数,那么,一方面难以得到机器学习期望的实际结构效果;另一方面,由于没有具体问题具体分析,正则化项的凸性和损失函数的光滑性没有得到充分的利用,将导致在理论上很难得到理想的收敛速度结果.

对于求解具体领域的优化问题,优化理论方面的著名学者 Nesterov 曾经指出:“黑箱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”[7].

本文考虑典型的稀疏学习优化问题,即 L1 正则化损失函数优化问题,以求解大规模问题为目的,介绍利用训练样本集合冗余和稀疏特点且保证正则化项结构的优化算法,其中包括在线优化、随机优化以及坐标优化方法的一些重要进展.

2 在线和随机优化方法

实际上,机器学习所研究的在线优化算法在优化理论中早已有所讨论,只不过使用的术语不同罢了.在优化

理论中,人们常常使用术语增量算法[8],其主要思路是:当目标函数由一些子函数之和组成时,可以通过每次仅对一个子函数进行“首尾相接”依次传递式的梯度优化迭代而最终得到原问题的最优解.可以证明,一般形式的投影次梯度方法就可以扩展成增量算法的形式,且具有和批处理方法同样阶的收敛速度[8].当我们按照随机的方式挑选子函数而不是按顺序依次进行优化时,此时的增量式梯度下降方法可以简单地从形式上称为随机梯度优化方法(stochastic gradient descent,简称 SGD).

正则化损失函数优化问题(1)具有子函数和的形式,且子函数具有明确的含义,即表示单个样本导致的正则化损失,因此,问题(1)的增量投影次梯度方法与在线学习算法的表现形式极为相似.而随机梯度下降算法可以表示为对随机挑选样本导致的损失进行优化,这和机器学习中随机优化的基本思想完全吻合.由于每次迭代计算仅仅是优化一个样本点造成的损失,与批处理算法相比大大缩减了内存开销.但也正是每次仅仅优化单个样本点造成的损失,收敛速度不可避免地会受到影响.然而,对机器学习问题来说,由于样本点是独立同分布的,当样本集合冗余时,达到一定泛化能力所需样本点往往只占大规模样本集合中很少甚至是极少的一部分,因此,只要针对部分样本点运行随机优化步骤后,优化问题解的学习精度就已经呈现出稳定的趋势,这就是随机优化方法特别适合求解大规模学习问题的主要原因.

虽然在形式上在线和随机优化差别甚微,似乎只是抽取样本方式上的差异,但近期的研究结果表明,它们在收敛性方面存在着重要的区别.下面围绕如何将不同一阶梯度优化算法扩展成为保持稀疏解的在线和随机优化算法,介绍近期的一些具有重要影响的研究成果.

2.1 一阶梯度在线方法

2003 年,Zinkevich 提出了一种在线优化算法[9].它实际上就是优化理论中增量投影次梯度方法处理学习问题的一种特例,即对于约束优化问题:

$$\min_{\boldsymbol{w}\in Q}\sum_{i=1}^{m} l_i(\boldsymbol{w}),$$

其中,Q 为闭凸集合,$l_i(\boldsymbol{w})=l_i(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$,文献[9]中,在线投影次梯度方法的主要迭代步骤为

$$\boldsymbol{w}^{t+1}=P_Q(\boldsymbol{w}^t-\eta^t\partial l_t(\boldsymbol{w}^t)),$$

其中,P_Q 为 Q 上的投影算子,η^t 为步长,∂l_t 为 $l_t(\boldsymbol{w})$的次梯度.为了衡量在线优化策略的优劣,定义 regret:

$$R(\boldsymbol{w})=\sum_{t=1}^{T} l_t(\boldsymbol{w}_t)-\sum_{t=1}^{T} l_t(\boldsymbol{w}).$$

从 regret 的定义中不难发现,它能够定量描述迭代过程中在线算法的解与批处理问题解之间的差异,从而可以定量描述在线优化算法的收敛速度.Zinkevich 的这项研究在机器学习优化算法的发展过程中具有极其重要的影响,它规范了机器学习对在线学习的认识,使得机器学习在线优化算法的研究有了统一的理论框架,摆脱了长期以来一直依赖启发式策略的尴尬.遗憾的是,上述研究仅仅局限于收敛性分析.或许是由于缺乏必要的实验结果,这项研究当时并未引起人们的充分关注.

在在线优化算法的基础上,很容易衍生出具有相应收敛速度的随机优化算法.2007 年,Shalev-Shwartz 首先对 L2 正则化的支持向量机问题使用强对偶定理得到优化问题的球形约束区域,从而使用随机投影次梯度在线算法进行求解,该方法称为 Pegasos[10].Pegasos 在大规模数据库上获得了很好的实验效果,处理 80 万个样本的 RCV1 数据库仅需几秒钟的时间.事实上,当对随机挑选的 5 000 个样本运行 Pegasos 以后,算法的学习精度已经趋于稳定.与当时的其他算法相比,Pegasos 的优势明显.至此,在线投影次梯度算法无论是在理论还是在应用上都取得了成功,引起了众多学者的广泛关注.由于优化问题(1)等价于约束问题:

$$\min_{\boldsymbol{w}\in Q}\sum_{i=1}^{m} l_i(\boldsymbol{w}),Q=\{\boldsymbol{w}:\|\boldsymbol{w}\|_1\leqslant\lambda_0\} \tag{2}$$

人们自然想拓广 Pegasos 求解优化问题(2).当 Duchi 于 2008 年提出了高效 L1 球投影算法以后[11],投影次梯度方法求解稀疏学习问题似乎迎刃而解.由于每一步迭代都进行 L1 投影运算,在线投影次梯度方法求解优化问题(2)很好地保证了解的稀疏性,但在每一步迭代中不可缺少的投影计算却是该方法难以克服的瓶颈,在实际运行

中会导致循环嵌套循环的问题.如何使投影计算更为简洁和高效,一直是人们关注的问题,一些文献采用了不使用投影或者仅仅在最后一步迭代中使用投影而中间过程不做投影运算的思路[12,13].另外,和批处理算法一样,在线算法对步长特别敏感,往往一开始能使目标函数急剧下降,但不久就会进入震荡或缓慢下降阶段,常用的方法是使用线性搜索[14].最近的研究进展表明,也可以采用求解对偶优化问题方法加以避免.

不难发现,问题(2)还是与 L1 正则化问题存在着一定的区别,人们还是想直接使用投影次梯度方法处理正则化问题.但当将整个正则化项和损失函数作为目标函数使用投影次梯度在线算法时,所得结果却与使用批处理形式的约束投影方法不同,这主要是由于浮点数计算不能抵消为 0,出现了解的稀疏性无法保证的现象,这意味着使用 L1 正则化项期望获得的稀疏性在优化过程中实际上并没有得到保证,此时,人们已经意识到在线和随机优化过程中保持正则化项结构的必要性.

在优化理论领域,除了典型的投影次梯度算法以外,还有很多高效的一阶梯度方法.自然地,能否将这些方法扩展成在线形式并保证正则化项的结构,是在线学习应该考虑的问题.2009 年,Xiao 将 Nesterov 的对偶平均优化算法[15]推广为在线形式,称为 RDA(regularized dual averaging)[16];随后,Duchi 等人于 2010 年将经典的镜面下降算法[17]拓展为在线形式,称为 COMID(composite objective mirror descent)[18],统一了在线算法研究很多零散的研究结果.为了能够处理正则化机器学习优化问题,RDA 和 COMID 分别对原有的优化方法进行了必要的改进.它们的共同点是:在优化过程中将正则化项和损失函数分别看待,对损失函数进行近似线性展开而保持正则化项不变.RDA 和 COMID 的区别仅仅是损失函数近似展开方式的不同,前者一阶项由所有迭代权向量的平均决定,后者由瞬时梯度决定.特别地,对 L1 正则化问题,对涉及正则化项和损失函数线性展开的优化子问题可以采用软阈值方法进行解析求解[16],即

$$\text{RDA: } \boldsymbol{w}^{t+1}=\arg\min_{\boldsymbol{w}}\left[\lambda\|\boldsymbol{w}\|_1+\frac{1}{t}\left\langle\partial\sum_{i=1}^{t}l(\boldsymbol{w}^i),\boldsymbol{w}\right\rangle+\frac{\beta_t}{t}\|\boldsymbol{w}\|^2\right],$$

$$\text{COMID: } \boldsymbol{w}^{t+1}=\arg\min_{\boldsymbol{w}}\left[\lambda\|\boldsymbol{w}\|_1+\langle\partial l(\boldsymbol{w}^t),\boldsymbol{w}-\boldsymbol{w}^t\rangle+\frac{\beta_t}{t}\|\boldsymbol{w}-\boldsymbol{w}^t\|^2\right].$$

其中,COMID 算法中的$\|\boldsymbol{w}-\boldsymbol{w}^t\|^2$可以换成更一般的 Bregman 项.当损失函数一般凸时,取步长β_t与$\sqrt{t}$同阶;当损失函数强凸时,取步长$\beta_t\leqslant\ln t$[16].与投影算子类似,软阈值方法可以有效保证每一步迭代中解的稀疏性.由于可以获得解析解,有效地避免了投影算子的计算代价问题.从上述在线算法的研究进展中不难看出,能否根据不同的优化原理得到一些新的在线算法,这项研究无论是在理论上还是在实际应用中都具有重要意义.

扩展拉格朗日交替方向乘子法 ADMM(alternating direction method of multipliers)是一种基于对偶空间的梯度下降方法[19,20],近几年来受到了机器学习领域的普遍关注,在图像分类等实际问题上有很好的表现[21,22].为了应用 ADMM 求解正则化问题,一般首先将问题(1)转化为如下等价形式的约束优化问题:

$$\lambda\|\boldsymbol{z}\|_1+\frac{1}{m}\sum_{i=1}^{m}l(\boldsymbol{w},(\boldsymbol{x}_i,y_i)),\ \text{s. t. } \boldsymbol{w}-\boldsymbol{z}=0.$$

对这种等价的优化问题,ADMM 的主要迭代步骤为

$$\begin{cases}\boldsymbol{w}^{t+1}=\arg\min_{\boldsymbol{w}}\left[\lambda\|\boldsymbol{w}\|_1+\langle\mu^t,\boldsymbol{w}-\boldsymbol{z}^t\rangle+\frac{\rho}{2}\|\boldsymbol{w}-\boldsymbol{z}^t\|^2\right]\\ \boldsymbol{z}^{t+1}=\arg\min_{\boldsymbol{z}}\left[\frac{1}{m}\sum_{i=1}^{m}l(\boldsymbol{z},(\boldsymbol{x}_i,y_i))+\langle\mu^t,\boldsymbol{w}^{t+1}-\boldsymbol{z}\rangle+\frac{\rho}{2}\|\boldsymbol{w}^{t+1}-\boldsymbol{z}\|^2\right],\\ \mu^{t+1}=\mu^t+\rho(\boldsymbol{w}^{t+1}-\boldsymbol{z}^{t+1})\end{cases}$$

其中,$\rho>0$ 为步长.值得指出的是,在对偶空间求解梯度是一件很复杂的事情,其计算代价甚至相当于求解原来的优化问题.在 ADMM 中,采取了对优化变量进行分解交替更新的方式,这也是交替方向名称由来的主要原因.目前,ADMM 在收敛速度上取得了一些重要进展,批处理 ADMM 具有关于变分不等式 $O(1/t)$的收敛速度[22].

为了处理大规模问题,人们已经将 ADMM 拓展为在线形式[23,24].和批处理 ADMM 一样,在线 ADMM 的交替更新也体现为对正则化项和损失函数项分别更新.由于仅仅涉及到单个样本的损失,这两个更新都可以获取

解析解.与 RDA 和 COMID 类似,使用软阈值方法可以保证在线 ADMM 迭代过程中解的稀疏性.

至此,在优化理论中占重要地位的多种一阶梯度算法已经被成功地扩展为在线优化的形式.对于目标函数为凸的情形,上述在线结构优化方法可以获得根号形式的 regret 界.而当目标函数凸性假设进一步增强为强凸情形时,上述在线结构优化算法都可以获得更好的对数形式 regret 界[25].在线优化方法的优点之一是与样本的数目无关,理论上可以处理任意规模的数据.有些研究者甚至认为,在线优化能够敏锐地捕捉到数据变化的趋势,进而可以解决数据非同分布和实时学习问题.但如何从统计学习理论的角度来研究参数漂移等实际问题,还存在着诸多困难.

2.2 一阶梯度随机方法

在优化领域,随机梯度方法实际上是在随机优化的框架下讨论的.随机梯度优化算法通常不要求知道目标函数的具体形式,仅需知道目标函数梯度的无偏估计即可.对于学习问题,在线和随机算法优化的目标也是不同的,即,在线优化算法一般应满足 regret 的平均值趋于 0,而随机优化算法的目标函数是

$$\lambda \| \boldsymbol{w} \|_1 + E_{(\boldsymbol{x}_i, y_i)} l(\boldsymbol{w}, (\boldsymbol{x}_i, y_i)).$$

在处理随机优化学习问题时,人们通常假设单个样本损失函数的梯度是整个训练集合上损失函数梯度的无偏估计.显然,如果样本点是独立同分布的,这个假设自然成立.正是在这个合理的假设条件下,处理学习问题的随机优化算法在形式上表现为每次仅优化随机抽取的一个样本导致的正则化损失.由于形式上的相似性,随机优化过程中正则化项的稀疏性问题可以采用与在线优化方法完全相同的技巧加以解决.

在比较实验中,训练样本集合的数目通常是固定的,一般都采取随机选取样本的方式来验证在线优化算法的性能,此时使用的算法实际上已经变为随机优化算法了.所以一段时间以来,机器学习随机优化算法的研究思路几乎都是首先提出在线算法,建立 regret 界,随后通过在线与随机算法之间的标准切换技巧,给出相应收敛速度的随机算法,Pegasos,RDA 和 COMID 都是按照这种方式进行的.但这种研究方式却在最优收敛速度方面遇到了问题.另外,人们往往会直观地认为随机算法的最优速度一定与批处理算法相当,只是差一个常数因子罢了.实际上,批处理和随机算法的最优收敛速度以及与在线算法 regret 界之间还是存在着一定的区别的,特别是对目标函数为强凸或者损失函数为光滑的情形.

对于目标函数是强凸的问题,在线算法 regret 界衍生的收敛速度只能达到 $O(\ln t/t)$,但通过在随机次梯度方法中嵌入内循环(epoch),可以获得 $O(1/t)$的收敛速度.据此,Hazan 等人指出,在线算法的分析方法并不适用于随机算法[26,27].实际上,在线算法 regret 定义的形式比较苛刻,而随机算法只讨论收敛速度.相对来说,收敛速度的分析较为宽松,因此对一些优化算法,有时需要避开在线优化 regret 界的分析而直接讨论随机算法的收敛速度.对优化算法 ADMM 来说,与其在线形式对应的收敛速度相比,近期的研究结果表明,随机 ADMM 获得了更好的收敛速度上界[23,24].

对于优化问题(1),当损失函数光滑时,即损失函数的梯度满足 Lipschitz 条件时,可以将损失函数进行局部二阶展开,从而近似地看成二次多项式,得到下述迭代算法:

$$\boldsymbol{w}^{t+1} = \arg\min_{\boldsymbol{w}} \left[\lambda \| \boldsymbol{w} \|_1 + \langle \partial l(\boldsymbol{w}^t), \boldsymbol{w} - \boldsymbol{w}^t \rangle + \frac{L}{2} \| \boldsymbol{w} - \boldsymbol{w}^t \|^2 \right],$$

其中,$\partial l(\boldsymbol{w}^t)$是 $\sum_{i=1}^{m} l_i(\boldsymbol{w})$ 的次梯度,L 是损失函数梯度的 Lipschitz 常数.此时,每一步迭代的目标函数具有可分离的性质,即每维可独立处理,因此可根据软阈值方法解析求解二阶展开式的最小值.基于这种迭代思想的批处理算法具有 $O(1/t)$的收敛速度[17].1983 年,Nesterov 采用按一定规律变化步长的插值技巧将这种算法的最优收敛速度加速为 $O(1/t^2)$[15,28].但由于随机算法中的随机因素的存在,整个训练集合上损失函数的梯度与单个样本损失函数的梯度之间不免存在着方差,这个方差的大小会影响到最优收敛速度.当损失函数满足光滑性条件时,无论目标函数是否强凸,随机和批处理算法的最优收敛速度均不在同一个数量级别[16].

从总体上说,一阶梯度随机方法简单、高效,并且与样本的数目无关.从已有收敛速度的研究结果来看,正则化损失函数机器学习优化问题中只需假设损失函数光滑性,对整个目标函数并没有要求,这与优化理论中黑箱

算法存在着明显的区别.另外,一些研究者还讨论了随机优化算法的泛化能力问题[29,30].至此,随机学习的研究无论是在优化方法还是在统计分析方面均取得了令人瞩目的进展.但是,一阶梯度随机优化算法应用于机器学习问题仍然存在着很多问题,特别是在稀疏学习优化中,由于迭代过程中采用平均权向量作为最终的近似解,这往往会导致稀疏程度比批处理算法得到的解要差.另外,对一般凸的非光滑损失函数情形,采取何种平均权向量的策略以获得随机算法的最优收敛速度问题也还没有完全解决.

3 坐标优化方法

在大规模数据处理方面,除了简单实用的 SGD 以外,另一个值得关注的方法就是坐标下降(coordinate descent,简称 CD).CD 的思路非常简单,主要是对高维优化问题采取各个击破的方法分而治之.具体的迭代步骤就是固定其他维坐标,一次仅对选中的一维坐标进行优化求解.坐标最小化方法可以追溯到优化领域的交替优化原理[31],其主要思路是:求出一维优化问题的解析解或近似解析解,进而根据目标函数的下降特性得到收敛性.其理论问题主要是:如何在不同凸性和光滑性的假设下,得到理想的收敛速度上界.

CD 以简洁的操作流程和快速的实际收敛效果,成为处理大规模优化问题,特别是稀疏文本数据的首选方法.在线算法和 CD 存在着明显的区别:首先从形式上看,前者每一步迭代需要解单个样本关于所有维数的优化问题,而后者每一步迭代需要解关于所有样本的单维优化问题;另外,在具体的执行过程中,SGD 一般不去计算整个训练集合上目标函数的值或者其梯度的值,因为计算一次需要遍历所有样本,这实际上已经与整个求解算法的计算代价相当,然而在 CD 中,每一步迭代都会计算训练集合上目标函数的值或梯度的值.因此,巧妙的计算技巧在 CD 中起着十分关键的作用.著名学者 Nesterov 甚至认为,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[32].

Paul Tseng 是 CD 方法研究方面的著名学者,他从优化理论的角度对多种问题建立了坐标优化算法,其中也包括了机器学习正则化损失函数的优化问题[33−35].但是,他的研究工作大都偏重于理论分析.机器学习的一些研究者分别针对具体的学习问题提出了一些实用的 CD,其中,对偶问题的置换循环 CD 是性能最好的一种.它的内循环遍历所有坐标,在进行外循环时对样本集合进行一次排序置换.如果在内循环中仅仅是随机抽取一维坐标进行优化,一般称为随机 CD.为清楚起见,下面分原问题和对偶问题的坐标优化方法进行讨论.

3.1 原问题的坐标优化方法

原始问题的坐标下降方法(primal coordinate descent,简称 PCD)是坐标下降方法中最常见的一种形式.对 L2 正则化项 L2 损失建立的坐标优化算法是 PCD 的典型例子[36].之所以能够对这类问题可以建立坐标优化算法,主要是因为损失函数具有很好的光滑性,单变量子问题可以近似求得解析解.文献[37]中的 PCD 正是利用广义导数对损失函数进行二阶近似展开求得解析解,并证明了这种循环坐标优化方法具有超线性的收敛速度.

损失函数二次展开项的系数对算法的实际性能往往会产生十分重要的影响,这种影响甚至超过了算法本身.为了进一步提高算法的效率,一般都采用线性搜索策略寻求合适的系数,这必然需要计算目标函数的值.另一方面,在求单坐标优化问题时还需要知道目标函数的梯度,因此在 PCD 的运行过程中,从目前常用损失函数的具体形式来看,权向量与所有样本之间的内积运算总是无法避免的,这种内积运算一般会导致高达 $O(mn)$的计算代价.如果每一次更新都需要这样复杂的计算代价,坐标优化方法在处理大规模数据方面将一无是处.然而,CD 涉及的内积运算具有一定的规律,在首次的更新中计算内积后,其后的更新无须重新计算,只需计算这两次更新的差异即可.坐标下降方法中求偏导数的技巧能够方便地计算这种差异,从而巧妙地避免了内积计算的瓶颈,使得更新一次的计算复杂度仅为 $O(m)$.

对于优化问题(1),当使用 L2 损失函数时,可以对目标函数使用 Lipschitz 条件展开,从而类似地建立 PCD,与 L2 正则化问题 PCD 的区别仅仅是求解单坐标优化子问题方式不同.由于此时使用的是 L1 正则化项,可以应用软阈值方法求取近似展式的解析解.但由于 L1 正则化仅仅具有较弱的凸性,循环坐标优化算法只是在较强的假设下得到了收敛速度[37].

不难发现,文献[37]中的 PCD 和文献[17]中的批处理算法都是建立在损失函数二次近似展开的基础上,但

批处理方法将目标函数变成可以分离形式的基础上同步独立处理各个坐标分量,而 PCD 是将每个坐标依次更新,即任一坐标的更新都是在所有前面坐标已经更新的基础上进行的.在这种思路的启发下,我们从批处理形式的结构优化方法如 RDA 和 COMID 出发来设计 PCD[38],突破了已有文献在解单坐标优化子问题需要损失函数光滑性的限制,可得到非光滑损失函数的坐标下降方法.根据批处理算法和随机 PCD 之间的关系,容易证明随机 PCD 和批处理算法具有同样阶的收敛速度.实验结果表明,所得到的非光滑损失函数随机 PCD 对正则化 Hinge 损失问题实现了坐标优化预期的效果.

从形式上看,PCD 仅仅以一种优化算法的身份出现在机器学习问题的求解中,保持正则化项结构的问题主要体现在子问题的求解上.即便如此,求解机器学习原问题的坐标下降方法仍然有很多问题没有解决.特别是一般情形下的循环坐标下降方法收敛速度,至今还是一个 Open 问题.

3.2 对偶问题的坐标优化方法

由于支持向量机的损失函数是非光滑的,在处理原问题时往往需要增添额外的辅助项来处理单维坐标优化子问题[16,38],但其对偶问题的目标函数却是二次多项式形式的.尽管存在着约束条件,对应的单维子问题是闭区间上的单变量二次优化问题,可以方便地解析求解.正是利用这一事实,研究者提出了对偶坐标下降方法(dual coordinate descent,简称 DCD)求解支持向量机问题[39],避开了原问题的非光滑性,并且得到了 DCD 关于对偶变量的指数收敛性.

与 PCD 完全类似,在求对偶问题单坐标优化子问题时也需要知道对偶目标函数的梯度,同样也可以采用增量的方式计算,将计算代价由 $O(mn)$减少至 $O(m)$,只不过这时对偶目标函数的偏导数可以使用大家熟知的权向量与支持向量的关系表示为权向量与样本的内积,从而可以借助于权向量的增量来高效计算.

由于 L2 正则化优化原问题是强凸的,因此其对偶问题目标函数具有光滑性[8].无论是对 SMO 批处理算法还是坐标优化算法,光滑性都给涉及的子优化问题求解带来了极大的便利,但对于 L1 正则化问题,由于其仅具有一般的凸性,尽管根据共轭函数可以得到对偶优化问题[8],但得到的对偶问题甚至比原问题还要难解.所以一段时间以来,很少有使用对偶方法求解 L1 正则化问题的报道.近期的研究表明,L1 正则化问题的求解可以转化为 L1 和 L2 混合正则化项的问题,这为使用 DCD 求解 L1 正则化问题铺平了道路,此时,可以使用多种方法近似解析求解单坐标子问题[40].对于光滑损失函数问题,当循环次数充分大时,DCD 关于原问题目标函数收敛速度的界比 SGD 目前收敛速度的界要好;而对于非光滑损失函数,DCD 关于原问题目标函数的收敛速度与 SGD 的最优收敛速度相当[39,41].这些理论分析保证了当 DCD 数次遍历训练样本集合以后,会获得较好的收敛效果.

从形式上看,在原问题中使用 SGD 随机抽取某个样本进行优化,相当于在对偶问题中使用随机 DCD 中求该向量对应坐标分量的系数,它们的更新方式非常相似,因此,随机 DCD 对于冗余样本问题也特别有效.但两者还是存在着一些区别:首先,SGD 不需要固定样本集合的大小,DCD 需要预先知道样本的个数并在迭代过程中存储所有样本对应坐标分量的系数;其次,它们具体的更新运算也是不同的,SGD 一般是人为地根据凸性条件设定衰减步长,而在 DCD 中,这种步长是通过使对偶问题最速下降近似解析求解得到的,特别当重复抽取某一样本时,SGD 无须考虑该样本先前的优化结果,而 DCD 需要在以前的结果上更新;另一方面,对于机器学习问题,直接处理原问题和求解对偶问题各有长处.因此,PCD 与 DCD 也各有特点[36].一般来说,相对于 DCD,PCD 更适合一些样本非零特征维数远小于样本个数的学习问题,是处理文本分类的首选方法[42].从应用方面来说,采用置换策略的循环 DCD 在大量高维数据库上取得了远比 PCD 和当前流行的一些算法更快的收敛效果.鉴于对偶坐标优化在解决大规模问题中如此重要的地位,林智仁教授继开发了著名的支持向量机软件包 LIBSVM 之后,又针对大规模数据的线性分类问题专门开发了 LIBLINEAR 软件包,其中包含了各种形式的坐标优化算法.

与简单、高效的 SGD 相比,DCD 不存在步长的选择问题,实际收敛性能往往更好.美中不足的是,由于更新方式的限制,在算法执行过程中需要存储每个损失函数对应的对偶变量坐标,这对于处理复杂问题,如 pairwise 形式的损失函数问题会带来难以承受的存储开销.另外,对于很多实际的大规模问题,坐标优化读取数据的时间远远高于优化算法执行所需要的时间,因此,分布式和并行处理策略成为必然的选择.但目前的分布式系统和并行框架并不是专门为机器学习优化算法设计的,如何充分利用机器学习优化问题的结构特点,采用高效分布式

或并行算法进行求解,是一个有前景的研究方向.

4 结束语

一阶随机梯度和坐标优化这两种方法在求解高维大样本问题时简单、高效,各有特点和优势.一阶随机梯度方法主要利用大规模数据独立同分布的统计规律;而坐标优化主要利用高维数据稀疏的特性,并具有复杂性较低的计算目标函数及其梯度的技巧.因此,坐标优化、随机和在线优化方法充分地利用了机器学习问题自身的特点,很好地满足了机器学习问题大规模和结构的实际需求.这两种优化方法已经成为机器学习主流的求解方法.

应该看到,目前大多数的大规模优化算法仅仅讨论是线性分类问题,并且处理的大都是稀疏或冗余数据问题,对大规模的稠密数据或者非线性分类优化问题的求解依然任重道远.

References:

[1] Vapnik VN. Statistical Learning Theory. New York: Wiley-Interscience, 1998.

[2] Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization. Annals of Statistics, 2004,32(1):56−85. [doi: 10.1214/aos/1079120130]

[3] Zhang T. Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research, 2004,5:1225−1251.

[4] Wang J, Tao Q. Machine learning: The state of the art. IEEE Intelligent Systems, 2008,23(6):49−55. [doi: 10.1109/MIS.2008.107]

[5] Bennett KP, Parrado-Hernández E. The interplay of optimization and machine learning research. Journal of Machine Learning Research, 2006,7:1265−1281.

[6] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society (Series B), 1996,58(1):267−288.

[7] Nesterov Y. Primal-Dual subgradient methods for convex problems. Mathematical Programming, 2009,120(1):221−259. [doi: 10.1007/s10107-007-0149-x]

[8] Bertsekas DP, Nedić A, Ozdaglar AE. Convex Analysis and Optimization. Belmont: Athena Scientific, 2003.

[9] Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent. In: Proc. of the Int'l Conf. on Machine Learning. 2003. 928−936.

[10] Shalev-Shwartz S, Singer Y, Srebro N. Pegasos: Primal estimated sub-gradient solver for SVM. In: Proc. of the Int'l Conf. on Machine Learning. 2007. 807−814. [doi: 10.1145/1273496.1273598]

[11] Duchi J, Shalev-Shwartz S, Singer Y. Efficient projections onto the l_1-ball for learning in high dimensions. In: Proc. of the Int'l Conf. on Machine Learning. 2008. 272−279. [doi: 10.1145/1390156.1390191]

[12] Hazan E, Kale S. Projection-Free online learning. In: Proc. of the 29th Int'l Conf. on Machine Learning (ICML 2012). 2012. 521−528.

[13] Mahdavi M, Yang T, Jin R, Zhu S, Yi J. Stochastic gradient descent with only one projection. In: Advances in Neural Information Processing Systems. 2012. 503−511.

[14] Tao Q, Sun Z, Kong K. Developing learning algorithms via optimized discretization of continuous dynamical systems. IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012,42(1):140−149. [doi: 10.1109/TSMCB.2011.2163506]

[15] Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 1983,27(2):372−376.

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

[17] Beck A, Teboulle M. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 2003,31(3):167−175. [doi: 10.1016/S0167-6377(02)00231-6]

[18] Duchi J, Shalev-Shwartz S, Singer Y, Tewari A. Composite objective mirror descent. In: Proc. of the 23rd Annual Workshop on Computational Learning Theory. ACM Press, 2010. 116−128.

[19] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 2011,3(1):1−122. [doi: 10.1561/2200000016]

[20] Tomioka R, Suzuki T, Sugiyama M. Super-Linear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation. Journal of Machine Learning Research, 2011,12:1537−1586.

[21] Yang AY, Sastry SS, Ganesh A, Ma Y. Fast L1-minimization algorithms and an application in robust face recognition: A review. In: Proc. of the 17th IEEE Int'l Conf. in Image Processing (ICIP). 2010. 1849−1852.

[22] He BS, Yuan XM. On the $O(1/n)$ convergence rate of the douglas-rachford alternating direction method. SIAM Journal on Numerical Analysis, 2012,50(2):700−709. [doi: 10.1137/110836936]

[23] Wang H, Banerjee A. Online alternating direction method. In: Proc. of the 29th Int'l Conf. on Machine Learning (ICML 2012). 2012. 1119−1126.

[24] Ouyang H, He N, Tran LQ, Gray A. Stochastic alternating direction method of multipliers. In: Proc. of the 30th Int'l Conf. on Machine Learning (ICML 2013). 2013. 80−88.

[25] Hazan E, Agarwal A, Kale S. Logarithmic regret algorithms for online convex optimization. Machine Learning, 2007,69(2): 169−192. [doi: 10.1007/s10994-007-5016-8]

[26] Hazan E, Kale S. Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization. Journal of Machine Learning Research-Proceedings Track, 2011,19:421−436.

[27] Rakhlin A, Shamir O, Sridharan K. Making gradient descent optimal for strongly convex stochastic optimization. In: Proc. of the 29th Int'l Conf. on Machine Learning (ICML 2012). 2012. 449−456.

[28] Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2009,2(1):183−202. [doi: 10.1137/080716542]

[29] Kakade SM, Tewari A. On the generalization ability of online strongly convex programming algorithms. In: Advances in Neural Information Processing Systems. 2008. 801−808.

[30] Rakhlin A, Sridharan K, Tewari A. Online learning: Beyond regret. arXiv preprint, arXiv:1011.3168v2, 2010.

[31] Bezdek JC, Hathaway RJ, Howard RE, Wilson CA, Windham MP. Local convergence analysis of a grouped variable version of coordinate descent. Journal of Optimization Theory and Applications, 1987,54(3):471−477. [doi: 10.1007/BF00940196]

[32] Nesterov Y. Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization, 2012, 22(2):341−362. [doi: 10.1137/100802001]

[33] Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications, 2001,109(3):475−494. [doi: 10.1023/A:1017501703105]

[34] Tseng P, Yun S. A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. Journal of Optimization Theory and Applications, 2009,140(3):513−535. [doi: 10.1007/s10957-008-9458-3]

[35] Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming, 2009, 117(1-2):387−423. [doi: 10.1007/s10107-007-0170-0]

[36] Chang KW, Hsieh CJ, Lin CJ. Coordinate descent method for large-scale L2-loss linear support vector machines. Journal of Machine Learning Research, 2008,9:1369−1398.

[37] Saha A, Tewari A. On the finite time convergence of cyclic coordinate descent methods. SIAM Journal of Optimization, 2013, 23(1):576−601. [doi: 10.1137/110840054]

[38] Tao Q, Kong K, Chu DJ, Wu GW. Stochastic coordinate descent methods for regularized smooth and nonsmooth losses. In: Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. LNCS 7523, Bristol, 2012. 537−552. [doi: 10.1007/978-3-642-33460-3_40]

[39] Hsieh CJ, Chang KW, Lin CJ, Keerthi SS, Sundararajan S. A dual coordinate descent method for large-scale linear SVM. In: Proc. of the 25th Int'l Conf. on Machine Learning. ACM Press, 2008. 408−415. [doi: 10.1145/1390156.1390208]

[40] Shalev-Shwartz S, Zhang T. Proximal Stochastic dual coordinate ascent. arXiv preprint, arXiv:1211.2717, 2012.

[41] Shalev-Shwartz S, Zhang T. Stochastic dual coordinate ascent methods for regularized loss minimization. Journal of Machine Learning Research, 2013,14:567−599.

[42] Yuan GX, Chang KW, Hsieh CJ, Lin CJ. A comparison of optimization methods and software for large-scale L1-regularized linear classification. Journal of Machine Learning Research, 2010,11:3183−3234.

陶卿(1965－),男,安徽长丰人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,模式识别,应用数学.
E-mail: taoqing@gmail.com

姜纪远(1989－),男,硕士生,主要研究领域为机器学习,模式识别.
E-mail: jyjianggle@gmail.com

高乾坤(1989－),男,硕士,主要研究领域为机器学习,模式识别.
E-mail: gaospeed@gmail.com

储德军(1978－),男,讲师,主要研究领域为模式识别,凸优化及其在机器学习中的应用.
E-mail: djun.chu@gmail.com