第37卷 第17期
Vol.37 No.17

计 算 机 工 程
Computer Engineering

2011年9月
September 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)17—0175—03 文献标识码：A 中图分类号：TP181

L1正则化机器学习问题求解分析

孔 康[a]，汪群山[b]，梁万路[a]

(解放军炮兵学院 a. 五系；b. 二系，合肥 230031)

摘 要：以稀疏学习为主线，从多阶段、多步骤优化思想的角度出发，对当前流行的L1正则化求解算法进行分类，比较基于次梯度的多步骤方法、基于坐标优化的多阶段方法，以及软L1正则化方法的收敛性能、时空复杂度和解的稀疏程度。分析表明，基于机器学习问题特殊结构的学习算法可以获得较好的稀疏性和较快的收敛速度。

关键词：L1正则化；机器学习；稀疏性；多阶段；多步骤

Solution Analysis of L1 Regularized Machine Learning Problem

KONG Kang[a], WANG Qun-shan[b], LIANG Wan-lu[a]

(a. No. 5 Department; b. No. 2 Department, New Star Research Institute of Applied Technology, Hefei 230031, China)

【Abstract】To deal with the new time and space challenges of the machine learning problem algorithms from large scale data, this paper focuses on sparse-learning and categorizes the L1 regularized problem's the-state-of-the-art solvers from the view of multi-stage and multi-step optimization schemes. It compares the algorithms' convergence properties, time and space cost and the sparsity of these solvers. The analysis shows that those algorithms sufficiently exploiting the machine learning problem's specific structure obtain better sparsity as well as faster convergence rate.

【Key words】L1 regularized; machine learning; sparsity; multi-stage; multi-step

DOI: 10.3969/j.issn.1000-3428.2011.17.059

1 概述

统计机器学习算法在理论和应用上[1]都取得了丰硕的成果。当前一般在“正则化项+损失函数”的框架下对正则化机器学习问题进行考察，这是对支持向量机(SVM)算法理论框架的有效拓展。

虽然L1正则化可以得出期望的稀疏性，但由于L1范数是不可导的，从而L1正则化问题很难得到对偶形式，传统的基于梯度的算法(最速下降法、牛顿法等)都无法对该问题加以解决；加之实际问题的规模越来越大，一些只能解决小规模问题的经典算法(如内点法)此时也无法直接使用了。正因为这些原因，海量数据的优化求解问题吸引了国内外众多优秀的工作组和知名学者的目光[2]。

传统的批处理(batch)算法(每次迭代遍历所有样本的信息，甚至需要处理海森矩阵)，由于无法突破计算机存储空间的瓶颈而面临无法回避的挑战。取而代之的随机(stochastic)和在线(online)算法是处理大规模数据的必然选择，并已经取得了事实上的成功。所谓随机算法，即每次迭代只从样本集中随机取一个样本来更新解向量。而在线算法则考虑样本是随时间序列逐个产生的，每次也只用一个样本更新解向量。如果能证明当样本数趋于无穷大时这2个算法与批处理算法之间的解的误差(即regret界)是趋于0的，则该随机/在线算法即是收敛的。同时，在保证算法收敛的前提下，收敛速度的快慢也是必须考虑的。

“多步骤(multi-step)”与“多阶段(multi-stage)”是当前机器学习界出现频率较高的词汇。对于迭代算法而言，所谓多步骤算法，即把问题的求解划分成若干规模很小的子问题来解决。单步计算代价小，但迭代次数多，如最速下降法采取衰减步长策略时，单步只须计算下降方向，计算代价小但收敛速度慢。多阶段算法的迭代次数少，但单步的计算代价大，如需要计算和存储海森矩阵的牛顿法等。若多阶段算法子问题能进一步细分成有限子问题，并且可以获得解析解(closed-form solution)，则多阶段算法即可能获得良好的性能。

当前，学术界一般认为，坐标下降方法(coordinate descent method)子问题可以精确求解，是成功的多阶段算法，收敛速度最快，可以达到$O(\log(1/\varepsilon))$。此外，对于多步骤算法，如有强凸性质保证，也可达到$O(1/\varepsilon)$的收敛速度，而对于一般凸问题，只能达到$O(1/\varepsilon^2)$，其中，ε是目标函数值与理论最优目标函数值之间的误差。

本文从多阶段、多步骤的优化角度出发，对L1正则化机器学习问题最新的研究进展进行分析，并对该问题未来的研究方向进行展望。

2 正则化机器学习问题的描述

对于独立同分布的训练样本集$S=\{(\boldsymbol{X}_1,y_1),(\boldsymbol{X}_2,y_2),\cdots,(\boldsymbol{X}_m,y_m)\}\in R^n\times\Psi$(二分类问题时，$\Psi=\{+1,-1\}$；回归问题时，$\Psi=R$)，正则化机器学习问题可以归结为求解下述优化问题：

$$\min_{\boldsymbol{W}} f(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_p^\sigma+\frac{1}{m}\sum_{(\boldsymbol{x},y)\in S} l(\boldsymbol{W};\boldsymbol{X},y) \tag{1}$$

其中，$\|\cdot\|_p^\sigma$表示Lp范数的σ次方；$\boldsymbol{W}\in R^n$；$l(\boldsymbol{W};\boldsymbol{X},y)$称为损失函数，它控制模型的训练精度；$\|\boldsymbol{W}\|_p^\sigma$称为正则化项，用于避免模型的过拟合；通过调整参数$\lambda$，可以得到兼有训练精度和泛化能力的模型。

当$p=0$时，L0范数正则化项$\|\boldsymbol{W}\|_0=|\{j:w_j\neq 0\}|$。此时的优化问题(式(1))即为稀疏学习问题。机器学习算法面临高

基金项目：国家自然科学基金资助项目“基于损失函数的统计机器学习算法及其应用研究”(60975040)

作者简介：孔 康(1982－)，男，硕士研究生，主研方向：模式识别，人工智能；汪群山，讲师；梁万路，硕士研究生

收稿日期：2011-03-24 **E-mail：**ln.kang.kong@gmail.com

维海量数据(如文本分类数据库达到10^7样本个数、10^9样本维数)的现实挑战，而稀疏学习得到的解向量非零维的数量会尽可能得少，这对于高维样本的特征选择具有重要意义。遗憾的是，L0 正则化的优化问题是 NP 完全的，目前无法有效求解。

当 $p=1$ 时，L1 正则化项为 $\|\boldsymbol{W}\|_1=\sum_{i=1}^{n}|w_j|$，它是 L0 正则化的一种很好的近似。从优化角度来说：(1)L1 正则化可以得到凸优化问题，使得问题的求解成为了可能；(2)如果问题的解确实具有稀疏性，在一定条件下，L1 正则化确实可以求解特征稀疏问题。事实上，L1 正则化的最小二乘回归问题(LASSO)最早是在信号处理领域提出的。其稀疏解对于信号/图像处理具有重要意义，可以利用少量的特征更好地表征带有噪声的信号，或使一个带有噪声的图像更加平滑。本文分析 L1 正则化的式(1)的求解，此时 σ=1。主要的损失函数有 hinge 损失、最小二乘、logistic 等损失函数。

3 算法分析

L1 正则化问题的求解研究方式方法很多，各有优劣。

3.1 基于次梯度的多步骤方法

目前，式(1)的多步骤解法多是(次)梯度下降法及其改进版本。次梯度(sub-gradient)的概念是梯度的有效推广，可以通过下式定义 $f(\boldsymbol{W})$ 在连续不可导点 $\boldsymbol{W}_0$ 处的次梯度：

$$\boldsymbol{g}=\{\boldsymbol{V}\mid\forall\boldsymbol{W}:f(\boldsymbol{W})\geqslant f(\boldsymbol{W}_0)+\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{W}-\boldsymbol{W}_0)\}$$

随机梯度法(SGD) $\boldsymbol{W}_{k+1}=\boldsymbol{W}_k-\alpha_k g(\boldsymbol{X}_i,\boldsymbol{W}_k)$ [3]，作为解决无约束凸二次优化的一般方法，属于多步骤算法，单步计算开销小，但收敛慢，没有好的停止准则。注意到，式(1)具有如下等价优化形式：

$$\min_{\boldsymbol{W}} f(\boldsymbol{W})=\frac{1}{m}\sum_{(\boldsymbol{x},y)\in S} l(\boldsymbol{W};\boldsymbol{X},y)$$
$$\text{s.t.}\quad \|\boldsymbol{W}\|_1\leqslant\rho \tag{2}$$

文献[4]给出了性能优越的 L1 球投影算子的求解算法。并对式(2)采用投影次梯度的方法加以解决 $\boldsymbol{W}_{k+1}=P_{L1}(\boldsymbol{W}_k-\alpha_k g(\boldsymbol{X}_i,\boldsymbol{W}_k))$。$P_{L1}(\cdot)$ 表示投影算子。简而言之，投影过程可以用下式表示(θ 是投影算子根据 ρ 和 $\boldsymbol{W}$ 计算得到的阈值)：

$$\boldsymbol{W}_{(j)}=\begin{cases}\max(\boldsymbol{W}_{(j)}-\theta,0) & \text{如果}\boldsymbol{W}_{(j)}\geqslant 0\\ \min(\boldsymbol{W}_{(j)}+\theta,0) & \text{如果}\boldsymbol{W}_{(j)}<0\end{cases} \tag{3}$$

L1 投影算子能保证解的稀疏性，并能有效提高收敛速度，但要付出 $O(k\log n)$ 的计算代价(k 为解向量非零元个数)，实现起来也很复杂。

文献[5]指出，式(1)的 SGD 方法在实际使用过程中并不具有稀疏性。为避免 L1 投影算子的计算复杂性问题，同时为确保在线算法解的稀疏性，文献[5]提出了截断梯度(Truncated Gradient)的方法，如式(4)所示：

$$\boldsymbol{W}_{(j)}=\begin{cases}\max(0,\boldsymbol{W}_{(j)}-\alpha\theta) & \text{如果}\boldsymbol{W}_{(j)}\in[0,\theta]\\ \min(0,\boldsymbol{W}_{(j)}+\alpha\theta) & \text{如果}\boldsymbol{W}_{(j)}\in[-\theta,0]\\ \boldsymbol{W}_{(j)} & \text{其他}\end{cases} \tag{4}$$

其中，$0<\alpha<1$；$\theta>0$；$j=1,2,\cdots,n$。该方法是一种为达到稀疏性目的而强制在梯度上进行的一种截断，并不具有明确的机器学习含义。尽管在形式上与式(3)很相像，并且文献[5]在一定条件下也证明了这种方法的 regret 界，但此时并不知道算法优化的目标函数是什么，这可能会涉及到复杂的非凸优化问题，同时也使得收敛速度难以分析。

文献[6]提出了 FOBOS 算法，对次梯度方法进行了重要的改进。该算法可用下式描述：

$$\boldsymbol{W}_{k+\frac{1}{2}}=\boldsymbol{W}_k-\alpha_k g(\boldsymbol{X}_i,\boldsymbol{W}_k)$$
$$\boldsymbol{W}_{k+1}=\arg\min_{\boldsymbol{W}}\left\{\frac{1}{2}\|\boldsymbol{W}-\boldsymbol{W}_{k+\frac{1}{2}}\|_2^2+\alpha_{k+1/2}\Psi(\boldsymbol{W})\right\}$$

其中，$\Psi(\boldsymbol{W})$ 可以取 L1、L2 以及 L∞ 正则化项。FOBOS 第 1 步是标准的 SGD 方法，第 2 步在最小化正则化项的同时保持尽可能靠近第 1 步的解向量。文献[6]进一步给出了算法的收敛性分析和 regret 界，从而给算法奠定了坚实的理论基础。当取 L1 正则化项时，第 2 步具有解析解，算法具有了稀疏性。

3.2 基于坐标优化的多阶段方法

坐标优化方法即坐标下降方法，其通过逐个优化解向量的每一维特征(坐标)，实现一次外循环。内循环中，在优化某一坐标时，固定 $\boldsymbol{W}$ 的其余 d-1 维坐标不动，对该维坐标(不妨设为第 j 维)求解如下单变量子问题：

$$\min_z D_j(z)\equiv f(\boldsymbol{W}+z\boldsymbol{e}_j)-f(\boldsymbol{W}) \tag{5}$$

其中，$\boldsymbol{e}_j=[\underbrace{0,0,\cdots,0}_{j-1},1,0,\cdots,0]^{\mathrm{T}}$。当 $f(\boldsymbol{W})$ 是可微函数时，式(5)具有解析解。

遗憾的是，式(1)的目标函数由于存在绝对值函数，是不可微的，致使式(5)不具有解析解，从而直接应用坐标下降方法成为了不可能。学者们提出了许多近似的方法[2]，如 Goodman 选择式(5)的上界 $A_j(z)\geqslant D_j(z)$，并对其进行优化；Genkin 提出 BBR 算法，将式(5)用类似于泰勒展开的方法进行变形，在信赖区域内对问题加以解决。但该算法的收敛性分析目前仍然无法给出。文献[2]对式(5)的损失函数进行二阶近似，此时得到解析解后再进行线搜操作，保证目标函数的单调下降性，从而保证了算法的超线性的收敛速度。

文献[7]将式(1)转化为如下等价问题：

$$\min_{\boldsymbol{W}^+,\boldsymbol{W}^-}\ \sum_{j=1}^{n}w_j^+ +\sum_{j=1}^{n}w_j^- +C\sum_{i=1}^{m}l(w^+,w^-;\hat{\boldsymbol{X}}_i,y_i)$$
$$\text{s.t. } w_j^+\geqslant 0,\ w_j^-\geqslant 0,\ j=1,2,\cdots,n \tag{6}$$

其中，$\hat{\boldsymbol{X}}_i=[\boldsymbol{X}_i;-\boldsymbol{X}_i]$。文献[7]仍用式(5)，但此时能够得到 $D_j^{''}(z)$ 的全局上界，而不需要在信赖域内解决问题，从而对单变量子问题的求解更直接有效，算法也有更加厚实的理论保证。

当然，式(6)所解决的问题将样本空间扩维一倍，对于解决海量问题来说，算法的时空效率仍值得商榷。

3.3 软 L1 正则化方法

文献[8]提出了正则化共轭平均(RDA)的方法，对式(1)的求解具有极其重要的意义。RDA 方法单步计算代价小，和多步骤算法相当，同时具有多阶段算法的一些优点。文献[8]指出，虽然数学优化方法的应用已经成为机器学习研究的核心内容之一，但两者是不能等同的。SGD 简单地将正则化项看做普通的凸函数，并用其次梯度方向来迭代解向量，这没有有效地发挥正则化项的作用，或者说没有真正发掘机器学习问题结构的特点。并且，像文献[5]这些基于 SGD 的改进算法也是不够可靠的，其稀疏性也不够理想。文献[8]将文献[4]的方法取名为“硬”L1 正则化，而将 RDA 看作是一种“软”的 L1 正则化解法。之所以这么命名，在于 RDA 将正则化项中引入了强凸辅助项 $h(\boldsymbol{W})=0.5\|\boldsymbol{W}\|_2^2$。其在线算法的迭代过程可以表述如下：

$$\boldsymbol{W}_{t+1}=\arg\min_{\boldsymbol{W}}\left\{\frac{1}{t}\sum_{\tau=1}^{t}\langle g_\tau,\boldsymbol{W}\rangle+\lambda\|\boldsymbol{W}\|_1+\frac{\beta_t}{t}h(\boldsymbol{W})\right\} \tag{7}$$

其中，$\{\beta_t\}$ 是非负、不减的序列；$\frac{1}{t}\sum_{\tau=1}^{t}g_\tau$ 表示所有次梯度(视为与解向量共轭的空间中的点)的均值，也是共轭平均名称的由来。

辅助项 $h(\boldsymbol{W})$ 对问题的解决带来了极大的便利：

(1)辅助项使得 RDA 算法获得了目前最优的收敛速率或 regret 界。即，对于 L2 正则化项可以获得 $O(\ln t)$ 的收敛速率；对于一般凸的 L1 正则化项问题可以获得 $O(1/\sqrt{t})$ 的收敛速率。

(2)辅助项使得式(7)的迭代可以获得解析解，这是区别于传统 SGD 方法的重要方面。也正因为解析解的获得，使得对解向量属性的归零更可靠更彻底，从而解的稀疏性从理论上保证了好于文献[5,7]。而文献[6]的 FOBOS 则是 RDA 方法的一个特例。以上“软”L1 正则化解法，其单步的计算代价与 SGD 相当，都是 $O(n)$，具有多步骤算法的计算代价小的特点；每一步解析解的获得又使得 RDA 具有多阶段算法快速收敛的特性。

实际上，式(7)为式(1)的解决提供了一般性的框架。不光解决了 L1 正则化的问题，也涵盖了 L2 正则化问题的求解。

4　结束语

本文分析了 L1 正则化机器学习问题的研究进展。可以看出，RDA 算法建立的框架，为式(1)的解决提供了更有效的方法。机器学习的优化过程和经典数学优化方法有很多不可忽视的差异。结构学习问题，即如何深入挖掘机器学习问题的结构特点，利用改进的数学优化手段解决实际问题，将是下一步的研究方向。

参考文献

[1] 黄诗华, 陈一民, 陆意骏, 等. 基于机器学习的自然特征匹配方法[J]. 计算机工程, 2010, 36(20): 182-184.

[2] Yuan Guoxun, Chang Kaiwei, Hsieh C J, et al. A Comparison of Optimization Methods for Large-scale L1-regularized Linear Classification[EB/OL]. (2009-11-04). http://www.csie.ntu.edu.tw/~cjlin/papers.html.

[3] Zhang Tong. Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms[C]//Proc. of the 21st International Conference on Machine Learning. [S. l.]: ACM Press, 2004: 919-936.

[4] Duchi J, Shalev-Shwartz S, Singer Y. Efficient Projections onto the L1-ball for Learning in High Dimensions[C]//Proc. of the 25th International Conference on Machine Learning. [S. l.]: ACM Press, 2008: 272-279.

[5] Langford J, Li Lihong, Zhang Tong. Sparse Online Learning via Truncated Gradient[EB/OL]. (2009-01-12). http://portal.acm.org/citation.cfm?id=1577097.

[6] Duchi J, Singer Y. Efficient Online and Batch Learning Using Forward Backward Splitting[EB/OL]. (2009-12-10). http://jmlr.csail.mit.edu/papers/v10/duchi09a.html.

[7] Shalev-Shwartz S, Tewari A. Stochastic Methods for L1 Regularized Loss Minimization[C]//Proc. of the 26th International Conference on Machine Learning. [S. l.]: ACM Press, 2009: 929-936.

[8] Xiao Lin. Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization[EB/OL]. (2010-10-11). http://jmlr.csail.mit.edu/papers/v11/xiao10a.html.

编辑　任吉慧

(上接第 174 页)

$$\omega=\tilde{\beta}^{+}=(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6)^{\mathrm{T}}=(0.20,0.10,0.25,0.10,0.15,0.20)^{\mathrm{T}}$$

利用文献[4]中的加权几何平均算子对表 1 中的医疗资源供应商隶属度矩阵进行计算，可以得到各供应商的综合区间直觉模糊值 $\tilde{A}_i(i=1,2,\cdots,5)$：

$\tilde{A}_1=\langle[0.497\ 2,0.619\ 0],[0.121\ 0,0.246\ 7]\rangle$

$\tilde{A}_2=\langle[0.439\ 8,0.567\ 3],[0.226\ 0,0.353\ 3]\rangle$

$\tilde{A}_3=\langle[0.520\ 4,0.630\ 7],[0.199\ 1,0.378\ 2]\rangle$

$\tilde{A}_4=\langle[0.325\ 7,0.484\ 8],[0.287\ 8,0.413\ 2]\rangle$

$\tilde{A}_5=\langle[0.489\ 6,0.633\ 8],[0.218\ 5,0.330\ 1]\rangle$

进一步利用文献[4]中的得分函数计算得到供应商 $\tilde{A}_i$ 的得分函数值 $\Delta(\tilde{A}_i)(i=1,2,\cdots,5)$ 为 $\Delta(\tilde{A}_1)=0.374\ 2$，$\Delta(\tilde{A}_2)=0.286\ 9$，$\Delta(\tilde{A}_3)=0.287\ 4$，$\Delta(\tilde{A}_4)=0.054\ 7$，$\Delta(\tilde{A}_5)=0.213\ 9$。

由于得分函数值越高说明该医疗资源供应商的素质越好，因此可根据 $\Delta(\tilde{A}_i)$ 的大小对最优供应商进行如下排序：$a_1>a_3>a_2>a_5>a_4$。可见最佳的医疗资源供应商同样为 a_1。通过对表 1 中医疗资源供应商的隶属度矩阵进行分析，可以看出上述结论是比较合理的。与其他算法相比，本文采用的医疗资源供应商选择模型和算法更实用。

6　结束语

本文针对不完全信息且准则值为区间直觉模糊集的多准则供应商排序问题，通过逻辑集成得到各供应商的区间直觉模糊集，并建立了各供应商的 Hamming 距离寻优模型，利用粒子群优化算法得到该模型的解。该方法满足了医疗救援组织在权重系数信息不完全条件下的模糊决策要求，为医疗资源供应商的选择提供了有益的决策参考。在医疗资源供应商的评价与选择过程中涉及的区间直觉模糊信息的集成及识别模式、模糊控制、模糊数据挖掘等其他领域中的应用还有待于研究和探索。

参考文献

[1] Shyur H J, Shih H S. A Hybrid MCDM Model for Strategic Vendor Selection[J]. Mathematical and Computer Modeling, 2006, 44(7/8): 749-761.

[2] Li Guodong, Yamaguchi D, Nagai M. A Grey-based Decision-making Approach to the Supplier Selection Problem[J]. Mathematical and Computer Modeling, 2007, 46(3/4): 573-581.

[3] 刘　杨, 胡仕成, 初佃辉, 等. 两阶段多供应商选择采购模型[J]. 计算机工程, 2009, 35(9): 74-76.

[4] 徐泽水. 区间直觉模糊信息的集成方法及其在决策中的应用[J]. 控制与决策, 2007, 22(2): 215-219.

[5] Boran F E, Genc S, Kurt M, et al. A Multi-criteria Intuitionistic Fuzzy Group Decision Making for Supplier Selection with TOPSIS Method[J]. Expert Systems with Applications, 2009, 36(8): 11363-11368.

[6] 王坚强. 一种信息不完全确定的多准则分类决策方法[J]. 控制与决策, 2006, 21(8): 863-867.

编辑　张正兴