文章编号：1007-1423(2015)32-0010-05　　DOI：10.3969/j.issn.1007-1423.2015.32.003

邻近点梯度法与交替方向乘子法求解 LASSO 的性能比较分析

陆萍

（苏州经贸职业技术学院机电与信息学院，苏州　215009）

摘要：

正则化模型是机器学习、压缩感知与推荐系统等领域的一类重要模型，其具有变量选择与稀疏化处理等功能，可以有效地避免模型的过拟合，完成信号重建或矩阵补全等工作。对稀疏正则化模型进行介绍，分析邻近点梯度算子与交替方向乘子法等最新的求解方法，并对它们的性能进行比较分析。

关键词：

正则化模型；LASSO；邻近点算法；交替方向乘子法

基金项目：

江苏省“青蓝工程”骨干教师培养对象，苏州经贸学院院科研课题 KY-ZR1407

0　引言

在机器学习与压缩感知等领域，为了获得具有更优泛化性能的模型，通常需要在求解模型时对最小化经验误差施加约束，从而达到模型选择的功能，以避免模型在训练集上取得优秀的性能，但在测试集上表现很差的情况。即通过对模型添加正则惩罚，避免发生模型“过拟合”现象。通过对模型添加正则化项，还可以达到增加唯一解的可能性与实现变量选择的功能，降低或避免仅使用经验风险最小化优化时带来的不适定问题，对模型起到修正作用，降低模型的复杂度。特别是在求解样本维度远高于样本数量的欠定方程中，适当的正则可以带来问题解的稀疏化，从而使得此类病态问题能够获得比较好的解。

在正则化项的选取上，岭回归[2](Ridge Regression)获得了广泛的应用，它不仅能够很好地降低模型的复杂度，避免模型的过拟合，而且使用的 L2 范数正则项具有光滑可导的优秀性质，可以使得模型在求解时直接获得解析解，在众多领域获得了广泛的应用。但 L2 范数正则不具备解的稀疏化能力，为此 Tibshirani[2]将其替换为 L1 范数，获得 LASSO(Least Absolute Shrinkage and Selection Operator)模型，产生稀疏模型的能力，而在取得稀疏解的同时亦即实现了变量的选择与降维，对于求解欠定问题非常有效。尽管在 LASSO 之前已有桥回归[1](Bridge Regression)模型，但在求解模型的算法上却不及求解 LASSO 的 LAR(Least Angel Regression，LAR)算法高效，因此未获得广泛的应用。与 LASSO 模型相类似，使用矩阵 Frobinus 范数、核范数、谱范数、迹范数等作为正则项的模型也在压缩感知、计算机视觉与推荐系统等领域中获得广泛的应用。

本文对使用 $L1$ 范数正则的 LASSO 模型进行了简要的介绍，并对最近提出的邻近点梯度方法[3](Proximal Gradient) 与交替方向乘子法 [4](Alternating Direction Multiplier Method，ADMM)两类适合于求解大规模问题的算法，在求解 LASSO 时的性能进行了比较分析。

1　LASSO 模型

对于线性回归模型：

$$y=w^Tx \tag{1}$$

其中 $x\in R^d$ 为回归变量，$w\in R^d$ 为权值向量，$y\in R$ 为对应的响应。若当前样本数为 N，可以通过 Least Square Regression 优化：

$$\min_w \frac{1}{2}\sum_{i=1}^{N}(y_i - w^T x_i)^2 \quad (2)$$

获得 w，使用矩阵表达为：

$$w=\arg\min_w \frac{1}{2}\|Xw-y\|_2^2$$
$$=(X^TX)^{-1}X^Ty \quad (3)$$

其中 $X\in R^{N\times d}$ 为样本矩阵，$y\in R^d$ 为响应向量。但由于抽取样本 x 中随机噪声的存在，所需解决的问题则成为：

$$y=X_w+\varepsilon \quad (4)$$

假设噪声变量为独立同分布，$E_i\sim N(0,\sigma^2)$。基于 Least Square 可获得解为 $w=(X^TX)^{-1}X^T(y-\varepsilon)$。然而当样本中数据的维数比较高时，使用 Least Square 会倾向于过拟合，一种有效的方法是选择尽量少的与输出相关度最高的变量维度，从而只使用这些维度进行回归，达到特征选择与降维的作用，而且可以比较好地解释数据，即获得与响应相关度最高的维度，此即为选择特征子集(Subset Selection)的思想。一种有效的选择方法即为最优子集选择（Best-Subset Selection），即从一个空的子集中逐渐加入与目标函数相关系数最大的特征，或从完整的特征集合中逐步丢弃掉相关度最低的特征。然而当样本的维度的相当高时，这种思想运算量过大。而采用 $L2$ 范数正则化的模型方法：

$$\min_w \frac{1}{2}\|Xw-y\|_2^2+\frac{\lambda}{2}\|w\|_2^2 \quad (5)$$

即岭回归对于回归系数虽然能够进行一定的压缩，但无法将其压缩为零，因此无法产生稀疏解，式中的 λ 为正则系数，其实现在对数据的拟合与正则之间的平衡。与之不同的是，如果将其中的 L2 范数替换为 L1 范数正则，则可以将较小的回归系数压缩为 0，从而可以产生稀疏解与实现特征选择：

$$\min w \frac{1}{2}\|Xw-y\|_2^2+\lambda\|w\|_1 \quad (6)$$

此即为 LASSO 模型。对于 LASSO 与岭回归的不同之处，在二维空间上如图 1 所示。左图为使用 $L1$ 范数正则的 LASSO 模型，右侧为使用 $L2$ 范数正则的岭回归模型。图中椭圆形显示的为风险误差函数的取值等高线，蓝色的菱形或圆形区域则对应于 $L1$ 与 $L2$ 范数正则项。由于 $L1$ 范数的约束，同时满足两者条件的点可取到部分维度为 0，但对于 $L2$ 范数由于其约束为圆形因此很难取得部分维度为 0 的解。

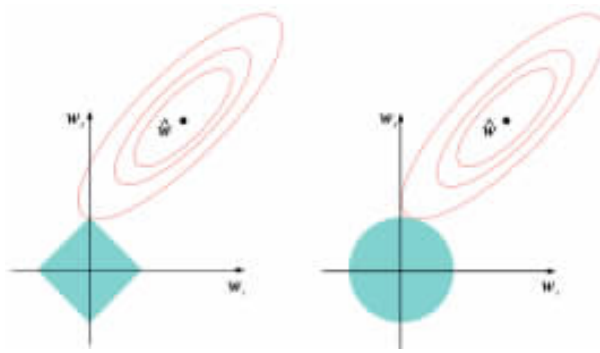

图 1 二维空间中 LASSO(左)与岭回归(右)示意图

2 邻近点梯度算法与交替方向乘子法

与岭回归具有显式解不同的是，由于 $L1$ 范数不可导，LASSO 无法获得其显式解，而只可以采用基于次梯度(Subgradient)的算法迭代求解。不过由于 LASSO 模型仍为凸函数，从而保证了算法的最优解的唯一性。在求解 LASSO 时，$L1$ 范数正则约束下的稀疏解在各维度组合上可以具有相当大的组合数，尤其是在样本维度高时，求解此问题成为 NP-hard 问题，直到 LAR 算法的提出，LASSO 才得以获得实际有效的应用。使用坐标下降(Coordinate Descent)类算法也可用来求解 LASSO 及其变形模型如 group LASSO，adaptive LASSO，sparse group LASSO 等问题。当前在凸优化领域基于邻近点算子（Proximal Operator）的邻近点梯度(Proximal Gradient Algorithm)算法，与基于分解思想的交替方向乘子法(ADMM)已被证明适合于求解大规模机器学习问题，它们也适用于求解 LASSO，这里对这两种算法进行性能比较与分析。

2.1 邻近梯度算法

首先定义函数 $f(x)$ 的邻近点算子为：

$$prox_f(v)=\arg\min_x f(x)+\frac{1}{2}\|x-v\|_2^2 \quad (7)$$

即为在当前点 $v\in R^d$ 的周围寻找极小化 $f(x)$ 的邻近点 x，因此可以通过设置初始点，通过不断迭代获得最优解 x^*。对于一般的使用非光滑范数正则的优化问题：

$$\min_x f(x)+g(x) \quad (8)$$

其中 $f(x)$ 为可微的凸函数，$g(x)$ 为任意的非光滑

不可微凸函数。邻近点梯度算法的迭代为：

$$x^{(k+1)}=prox(x^{(k)}-\eta k\cdot\nabla f(x^{(k)})) \tag{9}$$

对于使用 $L1$ 范数约束的 LASSO，由于 $L1$ 范数可以求得其次梯度，其迭代过程化为逐点运算的软阈值(Iterative Soft-thresholding Algorithm,ISTA)算法：

$$S_\lambda(x)=\begin{cases}x-\lambda, x>\lambda\\ 0, |x|\leqslant\lambda\\ x+\lambda, x<\lambda\end{cases} \tag{10}$$

基于邻近点梯度算法，在迭代求解时，不仅使用前一次搜索到的邻近点 $x^{(k)}$，还使用更前一次搜索到的点 $x^{(k-1)}$，即：

$$y^{k+1}=x^k+\alpha^k(x^k-x^{k-1})$$
$$x^{k+1}=prox(y^{k+1}-\eta_k\cdot\nabla f(y^{k+1}) \tag{11}$$

其中 α^k 取值为(0,1)且随迭代逐渐递减，可以达到加速效果，即获得加速邻近点梯度算法。一种简单的方法即为取 $\alpha^k=\frac{k}{k+3}$。

2.2 ADMM 方法

ADMM 算法基于对变量分解与坐标轮换的思想，对于形如：

$$\min_x f(x)+g(z)$$
$$s.t.\ Ax+Bz=c \tag{12}$$

的优化问题，创建如下的增广 Lagrange 目标函数：

$$L_\rho(x,y,z)=$$
$$f(x)+g(z)+y^T(Ax+Bz-c)+\frac{\rho}{2}\|Ax+Bz-c\|_2^2 \tag{13}$$

与式(8)类似，式(12)中 $f(x)$ 与 $g(z)$ 均为凸函数，通常 $f(x)$ 可微，而 $g(z)$ 不可微。其中为增广 Lagrange 系数。通过对此增广 Lagrange 函数中涉及的变量轮流优化即可获得最优解。其一般迭代框架为：

$$x^{k+1}=\arg\min_x f(x)+L_\rho(x,z_k,y_k)$$
$$z^{k+1}=\arg\min_z L_\rho(x^{k+1},z,y^k)$$
$$y^{k+1}=y^k+\rho(Ax^{k+1}+Bz^{k+1}-c) \tag{14}$$

若令 $r=Ax+Bz-c$，$u=\frac{\gamma}{\rho}$，则可获得 ADMM 的缩放形式(Scaled Form)：

$$x^{k+1}=\arg\min_x f(x)+\frac{\rho}{2}\|Ax+Bz^k-c+u^k\|_2^2$$
$$z^{k+1}=\arg\min_z g(z)+\frac{\rho}{2}\|Ax^{k+1}+Bz-c+uk\|_2^2$$
$$u^{k+1}=u^k+Ax^{k+1}+Bz^{k+1}-c \tag{15}$$

但与一般迭代算法不同，ADMM 算法在迭代收敛的停止准则上为双条件停止阈值判定，即原问题残差与对偶残差均要达到收敛阈值：

$$\|Ax^{k+1}+Bz^{k+1}-c\|_2^2\leqslant\xi^{pri}$$
$$\|\rho A^TB(z^{k+1}-z^k)\|_2^2\leqslant\xi^{dual} \tag{16}$$

3 邻近点梯度算法与 ADMM 算法的性能比较

为了对邻近点梯度算法与 ADMM 算法的求解 LASSO 的性能进行比较，在实验中选取样本维度为中等规模的 d=2500，为了进一步查看算法求解次定问题的性能，选择样本数为 N=500。样本各维度均由服从 N(0,1)分布的随机抽样获得，对回归系数 w 的稀疏度取为 0.05，且各元素服从 N(0,1)标准正态分布，并对正确响应向量添加 0.001 倍的高斯噪声。实验硬件环境为 Core i7 3720 CPU+8GB RAM，采用 MATLAB 环境，对邻近点梯度算法(PG)、加速邻近点梯度算法(APG)与 ADMM 算法的标准耗时与最优目标函数值进行了比较分析。实验结果如下：

表1 算法性能比较

	CVX	PG	APG	ADMM
目标函数值	21.9689	21.9747	21.9699	21.9012
耗时(s)	78.2853	0.7981s	0.4462	0.2789

表中的“CVX”为采用 CVX 优化工具箱直接求解结果。由表 1 可以看出 ADMM 算法在求解结果的性能上明显优于邻近点梯度算法及其加速版本，无论是在求解的目标函数值的精度上还是在算法的执行耗时上，其性能都非常突出，可见 ADMM 算法在求解问题时具有显著的优势。而对于邻近点算法较之于基本优化算法也具有相当不错的效果，在耗时上只需基本优化算法的 1%，而其加速版本中由于利用了再前一次的搜索到的邻近点信息，在求解精度上能够稍有改进，而耗时上也减少接近一半。

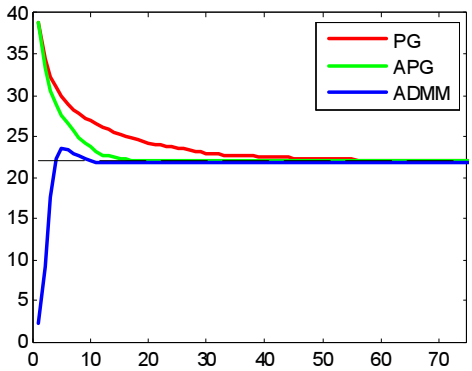

图 2 各算法目标函数值迭代曲线

上述各算法的目标函数值迭代曲线如图 2 所示。由图中可以看出 ADMM 的实际迭代次数也明显少于其他算法,能够很快收敛。

4 结语

本文对 LASSO 模型进行了介绍,对最近提出的邻近点梯度算法与交替方向乘子法在求解 LASSO 问题的框架进行了分析,并通过实验对两类算法在求解中等规模 LASSO 问题的性能上进行比较分析。实验结果表明交替方向乘子法无论在求解精度还是在算法耗时上都具有显著优势,因此也更适合于求解大规模机器学习问题。

参考文献:

[1]刘建伟,崔立鹏,刘泽宇,罗雄麟. 正则化稀疏模型综述[J].计算机学报,2015,38(7).

[2]Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning:Data Mining,Inference and Prediction[M]. New York: Springer Verlag, 2001.

[3]Neal Parikh, Stephen Boyd. Proximal Algorithms[J]. Foundations and Trends in Optimization,2013,1,(3):123-231.

[4]Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato and Jonathan Eckstein. Distributed Optimization and Statistical Learning Via the Alternating Directional Method of Multipliers[J]. Foundations and Trends in Optimization,2010,3(1):1-122.

[5]Ingrid Daubechies,Michel Defrise,Christine De Mol. An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint. arXiv, 2013.

[6]Nicholas G. Polson, James G. Scott. Proximal Algorithms in Statistics and Machine Learning, arXiv, 2015.

[7]Simon N,Friedman J,Tibshirani R. A Sparse-Group LASSO. Jounal of Computational and Graphical Statistics,2013,22(2):231-245

[8]李亚峰. 稀疏正则化的多目标图像分割变分模型[J].电子学报,2013,7(7):1329-1336

作者简介:

陆萍(1979-),女,江苏太仓人,讲师,硕士,研究方向为优化算法、图像处理

收稿日期:2015-11-05 修稿日期:2015-11-10

Performance Comparison and Analysis of Proximal Gradient and ADMM for Solving LASSO

LU Ping

(Suzhou Institute of Trade and Commerce, Suzhou 215009)

Abstract:
The regularized models play an important role in a lot of fields, such as: machine learning, compressing sensing, recommending system, and so on. With the ability of variable selection and generating sparse solution, the regularized models can avoid over-fitting. They may also be applied to signal reconstruction and matrix completion. Introduces the regularized models, and analyzes two recently developed algorithms: proximal gradient and ADMM, compares the performances on solving LASSO.

Keywords:
Regularized Model; LASSO; Proximal Gradient; ADMM

(上接第 9 页)

Multiple Attribute Decision Making Based Preference Ording

NA Di, WANG Li-qun, ZHANG Quan

(Shenyang University of Technology, Information Science and Engineering, Liaoning 110870)

Abstract:
Proposes an evaluation method based on preference order to determine the weight of expert in multi-attribute decision. Presents a method based on preference to determine the sequence of multi-attribute decision attribute weights. The preference order evaluation matrix converts to binary semantic after normalization, and then according to the comprehensive property rights to definition of the uncertainty and the degree of deviation, builds function model for determining the objective expert weight, gives a numerical example to prove problem description.

Keywords:
Tuple Linguistic; Multiple Attribute Decision-Making; Deviation; Uncertainty; Expert Weights