SUMMARY OF LASSO AND RELATIVE METHODS

Xia Jianan, Sun Dongyi and Xiao Fan
Department of Information and Computational Science, Beijing Jiaotong University, Beijing, China

Keywords: Feature selection, Variable selection, Pattern recognition, LASSO, Ridge regression.

Abstract: Feature Selection is one of the focuses in pattern recognition field. To select the most obvious features, there are some feature selection methods such as LASSO, Bridge Regression and so on. But all of them are limited in select feature. In this paper, a summary is listed. And also the advantages and limitations of every method are listed. By the end, an example of LASSO using in identification of Traditional Chinese Medicine is introduced to show how to use these methods to select the feature.

1 INTRODUCTION

The traditional linear regression model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_N x_N + \varepsilon \quad (1)$$

Where $(x_i, y_i), i = 1, 2, ...N,$ are data. $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ are regressors and y_i are response for the i th observation. The coefficients are what we need.

$$\beta = (\beta_1, \beta_2, ..., \beta_N) \quad (2)$$

According to the ordinary least square (OLS), we can minimize residual squared error

$$\sum_{i=1}^{N} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \quad (3)$$

and solve it to find the estimator of β which is expressed as

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_N) \quad (4)$$

and the solution by OLS is

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (5)$$

By using the OLS method, coefficients can be obtained, and so obtain the solution of the linear regression model. However there are some problems. The most important problem is that as the dimension get higher, the OLS method are no longer perfect, especially in predicting and selecting obvious coefficients. So based on the above statements, there are a new method to consummate the solution - feature selection. It's an effective way to choose the most useful coefficients from thousands upon thousands coefficients, in order to optimize the model.

Now we are going to discuss some of the method which can solve the feature selecting problem. In this paper, we list the advantages and disadvantages of every method and through the compare we give a summarization to the method of feature selecting. Also we show some of the methods which are not be solved completely yet. These problems are the next goal we want to solve.

2 ANALYSIS

2.1 Nonnegative Garrote

Breiman (1993) proposed the non-negative garotte, it starts with OLS estimates and shrinks them by non-negative factors whose sum is constrained. In these studies, Breiman proved the garotte has consistently lower prediction error than subset selection and is competitive with ridge regression except when the true model has many small non-zero coefficients. The garotte can get result in overfit or highly correlated settings. It can do a little than LASSO in the small number of large effects.

2.2 Previous Regression Method

2.2.1 Bridge Regression

Frank and Friedman (1993) proposed Bridge Regression which is a common form of penalized OLS.

Jianan X., Dongyi S. and Fan X..
SUMMARY OF LASSO AND RELATIVE METHODS.
DOI: 10.5220/0003426901310134
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 131-134
ISBN: 978-989-8425-54-6

$$\hat{\beta} = \arg\min\left\{\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2\right\} \quad subject\ to \sum_j |\beta_j|^q \le t \tag{6}$$

When $q = 1$ it is LASSO and when $q = 2$ it becomes Ridge Regression.

2.2.2 Ridge Regression

Hoerl and Kennard (1970) Proposed Ridge Regression, is the earliest use of punishment to achieve this purpose least square thinking.

$$\beta^{ridge} = \arg\min\left\{\sum_{i=1}^{N}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{N}\beta_j^2\right\} \tag{7}$$

Here is a control factor to reduce the extent of the parameters: the larger the value of income, the greater the degree of reduction.

2.3 LASSO

2.3.1 The Traditional LASSO

Define:

$$\hat{\beta} = \arg\min\left\{\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2\right\} \quad subject\ to \sum_j |\beta_j| \le t \tag{8}$$

The LASSO (Tibshirani, 1996) can solve the problem in the traditional method of the feature selection. For example, too many useless coefficients are existed in the traditional method. Also, the computation is large. By limit the coefficients, some of them can be zero and by doing this can we make the feature selection. LASSO variable selection has been shown to be consistent under certain conditions. With the research of LASSO, many of the corresponding algorithms have been proposed. One of the earliest algorithms is the "Shooting" algorithms proposed by Fu (1998).Then Osborne, M . R. proposed that the path of the regressive solution is piecewise linear and also proposed the "homotopy" algorithms. Bradley Efron (2004) proposed the LARS algorithms which expound the relationship between the LASSO and Boosting. This algorithm solves the problem in computation perfectly. Zhao and Yu (2007) proposed the Stagewise LASSO algorithm to solve LASSO and Boosting.

2.3.2 Fused LASSO

However, there still are limitations in LASSO when facing particular problems, so improving the method becomes the main idea of researchers. Tibshirani, R proposed Fused LASSO which is based on LASSO as they take the order between features in some meaning way into consideration. The improved method was effective in solving variable selection problem with order between features. The Fused LASSO penalized coefficients as well as the difference of neighboring coefficients.

$$\hat{\beta} = \arg\min\left\{\sum_{i=1}^{N}(y_i - \sum_j x_{ij}\beta_j)^2\right\} \quad subject\ to \sum_{j=1}^{p}|\beta_j| \le t_1 \quad and \sum_{j=2}^{p}|\beta_j - \beta_{j-1}| \le t_2 \tag{9}$$

The first constraint provides sparsity in coefficients and the second one gives sparsity in difference of coefficients. Tibshirani provide 2 solutions for Fused LASSO using in unordered problems:

a) Order the features using multidimensional scaling or hierarchical clustering.

b) Remark the index of features. The Fused LASSO just require the order of neighboring features instead of all features, so, for each j we can set the index $k(j)$ to the closet feature of feature j , and then the second constraint becomes $\sum_{j=2}^{p}|\beta_j - \beta_{k(j)}| \le t_2$

A drawback of Fused LASSO is the computational speed is quite slow, especially as $p > 2000$ and $N > 200$, the method is totally out of work. The method of solve the problem is still a topic.

2.3.3 Adaptive LASSO

Consider the Adaptive LASSO

$$\hat{\beta} = \arg\min_{\beta}\left\|y - \sum_{j=1}^{p} x_j\beta_j\right\|^2 + \lambda\sum_{j=1}^{p}\omega_j|\beta_j| \tag{10}$$

and ω is a known weights vector. Compare to the LASSO, Adaptive LASSO assign different weights to different coefficients, not forces the coefficients

to be equally penalized in the penalty. Hui Zou (2006) shows that the weighted LASSO will have the oracle properties if the weights are data-dependent. However LASSO variable selection can to be consistent under certain conditions. We can use the current efficient algorithms for solving the LASSO to compute the Adaptive LASSO estimates. Nicolai Meinshausen （2007） proposed Relaxed LASSO to solve the over-compression problem.

2.3.4 Relaxed LASSO

Definition:

$$\hat{\beta}^{\lambda,\varphi} = \arg\min N^{-1}\sum_{i=1}^{N}(Y_i - X_i^T\{\beta\cdot 1\mu_\lambda\})^2 + \varphi\lambda\|\beta\|_1 \quad (11)$$

$$\mu_\lambda = \{1 \le k \le p \,|\, \hat{\beta}_k^\lambda \ne 0\} \quad (12)$$

$$\{\beta\cdot 1\mu_\lambda\}_k = \begin{cases} 0, k \notin \mu_\lambda, \\ \beta_k, k \in \mu_\lambda. \end{cases} \quad (13)$$

It is obvious that in the Relaxed LASSO, λ is no longer the only penalty for the coefficients, but based on the LASSO added another penalty variable φ. When $\varphi = 1$, the Relaxed LASSO and the LASSO are identical; When $\varphi < 1$, Relaxed LASSO will show the more sparse solution compare with the LASSO. And especially when $\varphi = 0$, there will be a degenerate solution.

LASSO is an effective method but it has some disadvantages. One of them is that when data are high- dimensioned, the rate of convergence is very slow. Relaxed LASSO has a lower complexity and by using this, it can both very effective and have a high rate of convergence. And Relaxed LASSO will get a sparser solution than LASSO. Also, it's solution both soft-thresholding and hard-thresholding estimators.

2.3.5 Group LASSO

When the dimensionality exceeds the sample size, it cannot assume that the active set of groups is unique, Yuan & Lin (2006） extends the former in the sense that it finds solutions that are sparse on the level of groups of variables, which makes this method a good candidate for situations described above, for handling discrete variables in the model selection.

2.4 SCAD

Fan and Li (2001) proposed SCAD (smoothly clipped absolute deviation).

$$P_\lambda'(\theta) = \lambda\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta < \lambda)$$

$$a > 2, \theta > 0$$

The penalty function makes the larger one of θ as same as OLS solution (unbiased). What is more, the solution is continuous.

The solution of SCAD was given by Fan(1997).

$$\hat{\theta} = \begin{cases} \mathrm{sgn}(z)(|z| - \lambda)_+ & |z| \le 2\lambda \\ \{(a-1)z - \mathrm{sgn}(z)a\lambda\}/(a-2) & 2\lambda < |z| \le a\lambda \\ z & |z| > a\lambda \end{cases}$$

2.5 Elastic Net

The LASSO is not a good choice to solve problems with the following 3 characters.

a) $p > N$. Someone has proved that the number of non-zero coefficients that LASSO can provide at most is $\max\{p, N\}$. So in the $p > N$ situation, there are at most p variables can be selected, which seems to be not enough to represent all the features of model.

b) If a group of coefficients show high correlations, the LASSO will choose one from the group, however, it doesn't concern which one has been chosen.

c) While $p < N$, there are high correlations among x_i, Ridge Regression would be a better choice.

H. Zou and T. Hastie (2003) offered a method based on LASSO to solve variables selection problems in these situations, called Elastic Net.

We introduce Naïve Elastic Net first. The criterion was given as following

$$\hat{\beta} = \arg\min\left\{\sum_{j=1}^{N}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right\} \quad (14)$$

When $\lambda_1 = 0$ it becomes Ridge Regression when $\lambda_2 = 0$ it appear to be LASSO, while $0 < \lambda_2 < \lambda_1$ or $0 < \lambda_1 < \lambda_2$, the criterion shows both features of Ridge Regression and features of LASSO.

3 APPLICATION

The quality of Traditional Chinese medicine is uneven. So it's necessary to distinguish between low

quality and high quality medicine. To meet the requirement, we need to select the effective features from Traditional Chinese medicine fingerprint.

Because the fingerprint characteristics is suitable for the methods we said before.so we try to apply LASSO to identify the quality of Traditional Chinese medicine fingerprint.

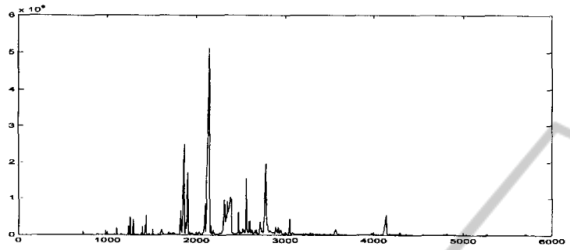

Figure 1: Traditional Chinese Medicine Fingerprint.

We treat each fingerprint as a observe (χ_i), the value of j th peak is value of j th feature (χ_{ij}), and y_j is symbol of medicine type identifier .
So Traditional Chinese Medicine Fingerprint Model:

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\}$$

$$subject \quad to \sum_j |\beta_j| \le t$$

The model has the same form as LASSO, and the solution of is also availiable for this model.

4 CONCLUSIONS

LASSO is a wonderful method for variable selection. However, nothing is perfect. To solve the weakness of LASSO, many techniques based on LASSO have been proposed. The fused LASSO is a most widely used method now, because it can meet the demands of many actual problems. But there is still no efficient computation method of Fused LASSO to solve complicate problems. Adaptive LASSO is a creative procedure which penalizes each coefficient with different weights. “Oracles Properties” is a good feature of Adaptive LASSO. Relaxed LASSO was raised to overcome the correlation of variables which has negative influence on predict accuracy of regression model. The solution of Group LASSO is sparse on the level of groups of variables. SCAD hold properties of sparse, continuous and unbiased. Elastic Net which holds advantages of both Ridge Regression and LASSO is another popular procedure.

REFERENCES

Breiman, L., 1995. *Better Subset Regression Using the Nonnegative Garrote*. Technometrics, 37, 373-384.

E. Frank and Jerome H. Friedman, 1993. *A Statistical View of Some Chemometrics Regression Tools.* Technometrics, 35(2), 109-135.

E. Hoerl, Robert W. Kennard., 1970. *Ridge Regression: Applications to Nonorthogonal Problems*. Technometrics, 12(1), 69-82.

Tibshirani, R., 1996. *Regression shrinkage and selection via the Lasso*. J. Roy. Stat. Soc. B, 58, 267–288.

Wenjiang J. Fu, 1998. *Penalized Regressions: The Bridge versus the Lasso.* Journal of Computational and Graphical Statistics, 7(3), 397-416.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R., 2004. *Least angle regression*. Ann. Stat., 32, 407–499.

Peng Zhao, Bin Yu., 2007. *Stagewise Lasso*. The Journal of Machine Learning Research, 8.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, Ji., Knight,Keith., 2005. *Sparsity and smoothness via the fused lasso*. Journal of the Royal Statistical Society Series B, 67, 91-108.

Zou, H., 2006. *The Adaptive Lasso and its Oracle Properties*. Journal of the American Statistical Association, 101, 1418-1429.

Meinshausen, N., 2007. *Relaxed Lasso Computational Statistics and Data Analysis*, 52,374-393.

Yuan, M., Lin, Y., 2006. *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49-67.

Zou, H., Hastie, T., 2005. *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society Series B, 67, 301-320.

Peng Xiaoling, 2005.V*ariable Selection Methods and Their Applications in Quantitative Structure — Property Relationship (QSPR).*

Yuan, M., Lin, Y., 2006. *Model selection and estimation in regression with Grouped variables*.Journal of the Royal Statistical Society Series B, 68, 49-67.

Fu, W. J., 1998. *Penalized regressions: the bridge VS the lasso*. Journal of Computational and Graphical Statistics, 7,397-416.