The Korean Journal of Applied Statistics
0000, Vol. 00, No. 0, 1–9

DOI: http://dx.doi.org/10.5351/KJAS.0000.00.0.001

Comparison study of shrinkage and selection methods for high-dimensional data

Junheui Lee

Department of Statistics, University of Korea

Abstract

In high-dimensional data analysis, we should beware of multi-collinearity between explanatory variables. If multi-collinearity exists between explanatory variables, each explanatory variable is affected by other explanatory variables, so the estimated coefficients have high variability. Therefore, problem arises in which explanatory variables important to the method may not be recognized. So, data analysts often choose shrinkage or selection method. In this paper, compare the performance of several shrinkage or selection methods through simulation.

Keywords: high-dimensional data, multi-collinearity, shrinkage method, selection method

1. 서론

Design Matrix에서 설명 변수의 개수가 많은 경우, 원하는 종속 변수를 설명하기 위한 모형을 적합시켰을 때, 모형이 너무 복잡해지는 문제가 발생할 수 있다. 이러한 대표적인 자료가 Amyotrophic Lateral Sclerosis (ALS) data이다. ALS data 분석의 목적은 369개의 연속적인 설명변수를 사용해서 ALS의 진행 속도 기능 등급 점수 (FRS)에 대한 예측을 하는 것이다. 해당 논문에서는 ALS data를 이용하여 설명 변수 중 일부를 사용해 적합시킨 여러 결과를 ensemble 하는 모형인 random forest (RF), gradient boosting (GB)와 모형의 coefficients 값을 축소시켜 중요한 설명 변수를 뽑아낼 수 있는 모형인 adaptive lasso (ALASSO), smoothly clipped absolute deviation (SCAD), elastic net (EN)을 사용할 것이다. 총 5가지 모형의 mean squared prediction error (MSPE)를 비교해 보고자 한다. 2장에서 EN에 대해 자세하게 설명한 후, 3장에서는 모의 실험을 진행하고, 4장에서는 3장의 결과를 바탕으로 결론을 내린다.

2. 모형 설명

2.1. Elastic net

Elastic net method (EN)는 L1 norm과 L2 norm을 동시에 사용하여 least absolute shrinkage and selection operator (LASSO) method와 ridge method의 장점을 모두 가지고 있어 변수의 숫자를 효과적으로 줄이고 싶을 때 사용하는 방법론이다 (Zou and Hastie (2005)). Ridge method는 설명 변수의 coefficients를 전반적으로 축소하고 LASSO method는 유의하지 않다고 판단한 설명변수의 coefficients를 0으로 만들게 된다. 다중 공선성이 존재할 경우, ridge method에서는 설명 변수들의 coefficients들이 축소되어 원활한 변수 선택이 어렵다는 문제가 발생할 수 있고, LASSO method에서는 실제로 유의한 설명 변수의 coefficient가 0이 되는 문제가 발생할 수 있다. 이러한 다중 공선성의 문제를 해결하고자 할 때, EN method 사용의 장점이 있다.

Footnote for research fund.

EN method의 coefficients 추정은 LASSO method와 ridge method의 L1, L2 norm을 동시에 고려한 손실 함수를 최소화 시키는 coefficients를 찾는 과정을 통해 이루어진다. EN method의 식은 아래와 같다.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{P}(\alpha \mid \beta_j \mid + (1-\alpha){\beta_j}^2)\},$$

EN method의 손실 함수는 기존의 OLS 손실 함수에 lagrange multiplier λ 를 이용해 L1, L2 norm 을 α를 이용해 가중 평균을 준 손실 함수를 더해준 것이다. α = 1인 경우는 EN method는 LASSO method와 동일하고, α = 0 인 경우는 EN method는 ridge method와 동일하다. EN method는 초모수 (hyper-parameter)인 λ의 값을 0 부터 ∞까지, α의 값을 0부터 1까지 조정해 최적의 coefficients 값을 찾는 것을 목적으로 한다.

3. 모의 실험

이 절에서는 모의 실험을 통해 1절에서 언급했던 5가지 method들의 성능을 비교하고자 한다. 모의 실험은 ALS data로 진행되었으며 각각의 method를 train set 1197개로 적합시켜, test set 625개에 대한 MSPE를 구하는 holdout 방식으로 진행했다. RF, GBM, ALASSO의 경우, 5회 bootstrap으로 bootstrap standard error of MSPE를 구할 수 있었으나, SCAD와 EN 같은 경우 seed의 변화에도 불구하고 값의 변화가 없어 bootstrap standard error of MSPE를 구할 수 없었다. 구체적인 과정은 각각의 소절에서 제시한다.

3.1. Random Forest

모의 실험에 사용한 RF method 의 초모수 (hyper-parameter)는 다음과 같다.

- `ntree`: Number of trees to grow.
- `mtry`: Number of variables randomly sampled as candidates at each split.
- `nodesize`: Minimum size of terminal nodes.

ntree는 200, 300, 400, 500을 사용하고, mtry는 총 설명 변수의 개수를 369개에 0.25,0.333,0.4를 곱한 값 92, 123, 148을 사용하고, nodesize는 2, 3, 5, 10을 사용해 48개 조합의 grid를 생성하여 초모수 조정을 진행했다. 그 중 가장 MSPE가 낮은 15개의 결과가 아래 table 1에 나타나 있다.

Table 1: Random Forest hyperparmeter tuning using 5 bootstrap

	num.trees	mtry	min.node.size	mspe	se.of.mspe
1	400	148	3	**0.2594**	0.000665
2	500	148	5	0.2596	0.000587
3	500	148	2	0.2597	0.000724
4	200	148	5	0.2601	0.000639
5	400	148	5	0.2601	0.000552
6	200	148	10	0.2602	0.001016
7	300	148	2	0.2603	0.000317
8	500	123	3	0.2603	0.000583
9	500	148	10	0.2603	0.000296
10	400	148	2	0.2604	0.000601
11	400	123	3	0.2604	0.000495
12	500	123	10	0.2604	0.000519
13	500	123	2	0.2606	0.000410
14	300	148	3	0.2606	0.000558
15	300	148	10	0.2606	0.000305

Table 1의 결과를 참고할 때 ntree = 400, mtry = 148, nodesize = 3 의 초모수를 사용할 때의 MSPE가 0.2594로 결과가 가장 좋다. 해당 초모수로 적합시킨 method의 variable importance plot이 아래 Figure 1에 나타나 있다.

Figure 1: *Variable Importance Plot of RF using* `ALS` *data.*

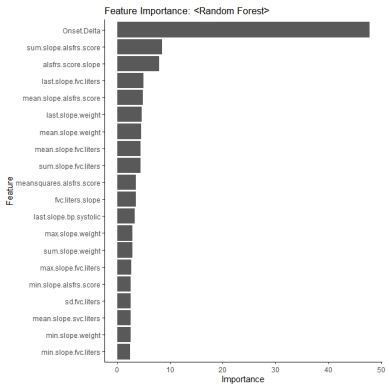


Figure 1의 결과를 참고할 때 Onset.Delta 설명 변수가 RF method의 node purity를 가장 크게 개선시키고 있고, 다음으로는 Symptom.Speech, Symptom.WEAKNESS 등의 변수가 node purity의 개선에 일조하고 있는 것을 알 수 있다.

3.2. Gradient Boosting

모의 실험에 사용한 GBM method 의 초모수는 다음과 같다.

- `shrinkage`: a shrinkage parameter applied to each tree in the expansion.
- `n.trees`: Integer specifying the total number of trees to fit.
- `interaction.depth`: Integer specifying the maximum depth of each tree.
- `n.minobsinnode`: Integer specifying the minimum number of observations in the terminal nodes of the trees.

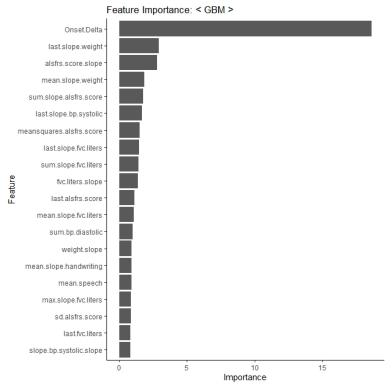
GBM method의 초모수 조정의 모의 실험은 먼저 나머지 초모수를 default 값으로 고정하고, shrinkage의 조정을 먼저 진행했다. shrinkage는 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1 을 사용했고, 최종적으로 shrinkage = 0.02를 선택했다. shrinkage를 0.2로 고정시킨 후에 n.trees는 200, 300, 400, 500을 사용하고, interaction.depth는 1,4,7을 사용하고, n.minobsinnode는 5, 10, 15를 사용해 36개 조합의 grid를 생성하여 초모수 조정을 진행했다. 그 중 가장 MSPE가 낮은 15개의 결과가 아래 table 2에 나타나 있다.

Table 2: GBM hyperparmeter tuning using 5 bootstrap

	n.trees	interaction.depth	n.minobsinnode	mspe	se.of.mspe
1	500	7	15	**0.2547**	0.000461
2	500	4	15	0.2552	0.000419
3	400	7	15	0.2558	0.000270
4	400	7	10	0.2560	0.000495
5	300	7	10	0.2564	0.000311
6	500	4	10	0.2569	0.000365
7	500	7	10	0.2569	0.000554
8	400	4	15	0.2572	0.000584
9	300	7	15	0.2578	0.000663
10	300	7	5	0.2587	0.000458
11	400	4	10	0.2587	0.000381
12	400	7	5	0.2588	0.000650
13	300	4	15	0.2588	0.000325
14	500	7	5	0.2595	0.000433
15	300	4	10	0.2595	0.000380

Table 2의 결과를 참고할 때 num.trees = 500, interaction.depth = 7, n.minobsinnode = 15 의 초모수를 사용할 때 MSPE가 0.2547로 결과가 가장 좋다. 해당 초모수로 적합시킨 method의 variable importance plot이 아래 Figure 2에 나타나 있다.

Figure 2: *Variable Importance Plot of GBM using* `ALS` *data.*



Rel.inf 는 각 설명 변수가 자식 노드로 분할할 때에 MSE를 감소시키는 정도를 나타내고, 상대적 영향력의 값이 클수록 감소량이 증가하게 된다. Figure 2의 결과를 참고할 때 변수 중요도 (rel.inf)가 제일 큰 설명 변수는 randomForest의 결과와 동일하게 Onset.Delta이다. 다음으로는 last.slope.weight, alsfrs.score.slope가 MSE의 감소에 크게 일조하고 있음을 알 수 있다.

3.3. Adaptive LASSO

모의 실험에 사용한 ALASSO method 의 초모수 는 다음과 같다.

- `lambda`: set optimal lambda leads to better convergence.

Test set을 사용해서 MSPE를 구하기 이전에, 10 fold cross-validation을 train set을 적용해서 0 과 100 사이의 lambda 값 중에 train MSE가 낮은 lambda 값 15개를 추출했다. 해당 lambda 값 15개에 대한 10 fold cross-validation의 결과가 아래 table 3에 나타나 있다.

Table 3: ALASSO 10 fold cross-validation hyperparmeter tuning

	lambda	MSPE	se.of.MSPE
1	0.123081	**0.2770**	0.015704
2	0.135081	0.2770	0.015562
3	0.112147	0.2771	0.015939
4	0.148251	0.2774	0.015453
5	0.102184	0.2774	0.016213
6	0.070431	0.2776	0.015912
7	0.064174	0.2776	0.015684
8	0.058473	0.2777	0.015458
9	0.093106	0.2778	0.016443
10	0.077298	0.2779	0.016295
11	0.162706	0.2779	0.015293
12	0.084835	0.2781	0.016615
13	0.053279	0.2782	0.015345
14	0.048546	0.2790	0.015307
15	0.178569	0.2790	0.015135

Table 3에서 얻은 train set에 대한 10 fold cross-validation으로 얻은 lambda 값 15개로 test set에 대한 MSPE를 구한 결과가 table 4에 나타나 있다.

Table 4: ALASSO hyperparmeter tuning for test MSPE using 5 bootstrap

	lambda	mspe	se.of.mspe
1	0.084835	**0.3806**	0.000165
2	0.102184	0.3808	0.000117
3	0.093106	0.3810	0.000477
4	0.077298	0.3813	0.000265
5	0.070431	0.3817	0.000219
6	0.112147	0.3819	0.000374
7	0.064174	0.3826	0.000122
8	0.123081	0.3838	0.000450
9	0.048546	0.3852	0.000286
10	0.053279	0.3853	0.000505
11	0.135081	0.3857	0.000339
12	0.148251	0.3887	0.000186
13	0.162706	0.3917	0.000623
14	0.178569	0.3942	0.000182
15	0.058473	0.3975	0.013486

Table 4의 결과를 참고할 때 lambda = 0.084835일 때의 MSPE가 0.3806로 결과가 가장 좋다. 또한, 최종 적합 모델의 유의한 변수는 133개로 나타난다.

3.4. Smoothly Clipped Absolute Deviation

모의 실험에 사용한 SCAD method 의 초모수는 다음과 같다.

- `lambda`: set optimal lambda leads to better convergence.

0 과 0.2 사이의 lambda 100개를 grid로 만들어서 SCAD method의 초모수 조정을 진행했다. 가장 MSPE가 낮은 lambda 값 15개에 대한 결과가 table 5에 나타나 있다.

Table 5: SCAD lambda hyperparmeter tuning using 5 bootstrap

	lambda	mspe
1	0.045770	**0.2682**
2	0.042690	0.2682
3	0.049080	0.2683
4	0.039810	0.2684
5	0.030120	0.2685
6	0.037130	0.2686
7	0.034630	0.2686
8	0.052630	0.2687
9	0.032290	0.2687
10	0.028090	0.2688
11	0.056430	0.2693
12	0.060510	0.2704
13	0.064880	0.2720
14	0.026190	0.2727
15	0.069570	0.2739

Table 5의 결과를 참고할 때 lambda = 0.045770를 사용할 때 MSPE가 0.2682로 결과가 가장 좋다. 또한, 최종 적합 모델의 유의한 변수는 122개로 나타난다.

3.5. Elastic Net

모의 실험에 사용한 EN method 의 초모수는 다음과 같다.

- `alpha`: The elastic net mixing parameter, alpha=1 is the lasso penalty, and alpha=0 the ridge penalty.
- `lambda`: set optimal lambda leads to better convergence.

Alpha는 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 를 사용하고, lambda는 0.05, 0.1, 0.15, 0.2, 0.25, 0.3 를 사용해 54개 조합의 grid를 생성하여 초모수 조정을 진행했다. 그 중 가장 MSPE가 낮은 15개의 결과가 아래 table 6에 나타나 있다.

Table 6: EN lambda hyperparmeter tuning using 5 bootstrap

	alpha	lambda	mspe
1	0.6	0.05	**0.2689**
2	0.5	0.05	0.2697
3	0.3	0.10	0.2700
4	0.7	0.05	0.2703
5	0.8	0.05	0.2707
6	0.9	0.05	0.2709
7	0.2	0.15	0.2711
8	0.4	0.10	0.2720
9	0.5	0.10	0.2731
10	0.1	0.25	0.2731
11	0.2	0.10	0.2733
12	0.4	0.05	0.2734
13	0.3	0.15	0.2739
14	0.1	0.20	0.2739
15	0.2	0.20	0.2746

Table 6의 결과를 참고할 때 alpha = 0.6, lambda = 0.05를 사용할 때, MSPE가 0.2689로 결과가 가장 좋다. 또한, 최종 적합 모델의 유의한 변수는 31개로 나타난다.

4. 결론

본 논문에서는 high-dimensional data인 ALS data에 다양한 shrinkage and selection method를 적용했다. 그 최종 결과는 table 7에 나타나 있다.

Table 7: result of hyperparameter tuning of shrinkage and selection methods

Method	MSPE
RF	0.2594
GBM	**0.2547**
ALASSO	0.3806
SCAD	0.2682
EN	0.2689

Table 7을 참고하면, GBM의 MSPE가 0.2547로 5개의 methods 중 가장 좋은 성능을 보인다. 특히, ALASSO method에 비해서 약 33%의 MSPE reduction이 있다.

한편, 각각의 모델에서 사용되는 설명 변수의 개수는 table 8에 나타나 있다.

Table 8: number of coefficients of shrinkage and selection methods

Method	number of coefficients
RF	148
GBM	185
ALASSO	133
SCAD	122
EN	31

Law of parsimony에 의해 비슷한 MSPE를 가진 method라면, 간단한 모형이 선호된다. table 7과 table 8을 참고했을 때 EN은 MSPE가 제일 낮은 모형인 GBM과 유사한 MSPE를 가지고 있고, 설명 변수의 개수가 31개로 다른 method들에 비해 작기 때문에 method의 가벼움 측면에서 다른 method들에 비해 뛰어나다.

References

Hui Zou and Trevor Hastie (2005). Regularization and variable selection via the Elastic Net, *Journal of the Royal Statistical Society, Series B*

Received 0, 0000; Revised 0, 0000; Accepted 0, 0000

고차원 데이터에 대한 축소 및 선택 모형 비교 연구

Junheui Lee

Department of Statistics, University of Korea

요 약

고차원 데이터 분석에서는 설명 변수 간의 다중 공선성에 유의해야 한다. 설명 변수간에 다중 공선성이 존재하는 경우, 각 설명 변수는 다른 설명 변수의 영향을 받기 때문에 추정된 계수의 변동성이 높다. 그렇기 때문에 중요한 설명 변수를 인식하지 못하는 문제가 발생할 수 있다. 이러한 문제점 해결을 위해 데이터 분석가들은 고차원 데이터 분석에서 축소 및 선택 모형을 선택할 때가 많다. 해당 논문에서는 시뮬레이션을 통해 다양한 수축 및 선택 방법의 성능을 비교하고자 한다.

주요용어: 고차원 데이터, 다중 공선성, 축소 모형, 수축 모형