Optimally adjusted last cluster for prediction based on balancing the bias and variance by bootstrapping

Jeong-Woo Kim*
*Dept. of Economics, Gangneung-Wonju National University
e-mail:kurtkim@gwnu.ac.kr

부트스트랩으로 조정되어 예측에 최적화된 마지막 군집

김정우*
*강릉원주대학교 경제학과

Abstract

Estimating a predictive model from a dataset is best initiated with an unbiased estimator. However, since the unbiased estimator is unknown in general, the problem of the bias-variance tradeoff is raised. Aside from searching for an unbiased estimator, the convenient approach to the problem of the bias-variance tradeoff may be to use the clustering method. Within a cluster whose size is smaller than the whole sample, we would expect the simple form of the estimator for prediction to avoid the overfitting problem. In this paper, we propose a new method to find the optimal cluster for prediction. Based on the previous literature, this cluster is considered to exist somewhere between the whole dataset and the typical cluster determined by partitioning data. To obtain a reliable cluster size, we use the bootstrap method in this paper. Additionally, through experiments with simulated and real-world data, we show that the prediction error can be reduced by applying this new method. We believe that our proposed method will be useful in many applications using a clustering algorithm for a stable prediction performance.

1. Introduction

When analyzing relationships among variables, including random variables with an equation of $y = f(x) + e$, where e is a random vector of a certain distribution with zero mean and preset covariance, two main problems are generally encountered: one is related to $f(x)$ and the other to e. The latter mainly includes heteroscedasticity and serial correlation problems, which induce efficiency of the estimator to decrease, thereby hindering rigorous hypothesis testing. The former, which is related to the misspecification problem leading to biased estimator, frequently presents a critical problem. In numerous applications, unless an underlying theoretical foundation for an approximation model has been rigorously established or is widely acknowledged in a related area, one is prone to obtaining a biased estimator.

Using a biased approximation model, the more instances we use, the greater the predictor bias we obtain. Therefore, with a biased approximation model, a subsampling approach could yield a more accurate performance up to the point where the gain from the reduced bias exceeds the loss from the increased variance [1]. Conversely, the subsampling approach could have serious drawbacks when the increased variance considerably exceeds the reduced bias because variance generally is inversely proportional to the sample size.

Moreover, searching for every neighbor reference of a target point, such as the k-nearest neighbor and moving average methods, is computationally expensive. Considering these factors, a more efficient method would be to partition the data into clusters.

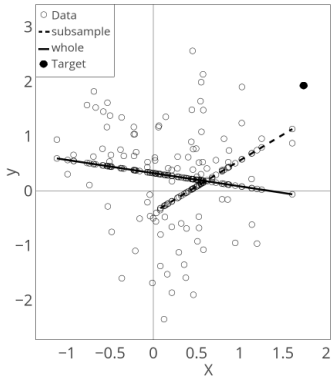
Clustering methods are easy to implement and extensively used as one of the unsupervised learning methods. However, in general, the number of clusters is not analytically obtained, but numerically searched in a set of candidate numbers. Thus, determining how to adequately adjust the size of the cluster would be a key approach to overcome the high variance and the overfitting problem. In particular, to adjust the size of the cluster for prediction, we may consider the cluster close to a prediction target, unlike the usual clustering method fitting a whole dataset. In the current study, the term *last cluster* indicates the cluster located closest to a prediction target after partitioning the data into clusters using a method such as k-means clustering because

considering only the last cluster in a prediction problem can be an efficient way to achieve accurate prediction.

The bootstrap method is well-known for reducing the bias [2] and variance [3] of an estimator. This method allows us to approximate an underlying functional form of a given dataset by averaging the noise of different bootstrapped samples out, thereby reducing bias. Also, the bootstrap method would be useful in balancing bias and variance, thus improving model prediction accuracy. Additionally, because a data-fitting-oriented model often leads to overfitting in many applications, it would be more efficient to adjust the size of the last cluster rather than to fit the entire dataset. In this study, the size-adjusted last cluster is termed as an adjusted last cluster (aCL), which has a size that falls between the size of the entire dataset and that of a last cluster by partitioning data.

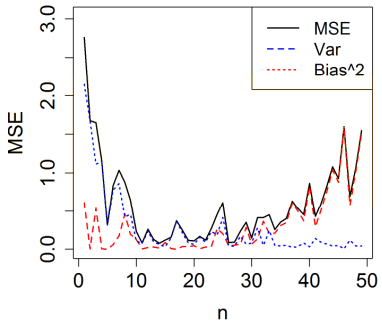
2. Methods

A linear estimator applied to the whole dataset depends on the too strong assumption; hence, this predictor produces a vulnerable prediction performance depending on the shape of the true function. Consider a linear estimator applied to a dataset generated from a quadratic functional form, the value predicted by this estimator over the whole dataset would be quite distant from the prediction target as shown in Fig. 1. However, the predicted value from a subsample of the given dataset achieves a more accurate prediction.



[Fig. 1] Prediction when using the whole dataset (whole, solid line) and a subsample (subsample, dashed line).

Furthermore, as more samples are used, the bias of the predictor increases but its variance decreases, as shown in Fig 2. In general, an estimator from a subsample suffers from high variance, whereas the estimator from a whole dataset suffers from high bias. Therefore, a size-adjusted subsample optimal for prediction should exist based on the balance between bias and variance [4]. For example, the size of such a subsample would be approximately 26, as seen in Fig 2.



[Fig. 2] Change of MSE (Mean squared error) according to the number (n) of samples used in estimation.

To obtain the optimal subsample for prediction, first it is necessary to partition a given dataset using a method such as k-means clustering. However, determining the size of the last cluster for prediction may not be easy. To reliably determine the size of the last cluster, knowing the change points, structural breaks and local extrema of the given DGP (Data Generating Process) can be useful. In practice, such values are not known a priori; one must guess the size of the last cluster by simply observing a graph of the DGP in an exogenous manner. However, partitioning the dataset in such a manner might be unreliable because the dataset observed is only a single realized sample among all the possible samples generated by the DGP. Thus, to obtain the size of the last cluster reliably, we would need to consider as many realized samples as possible, which would be unfeasible in practice. As an alternative, the bootstrap method used in [5] can be useful for virtually mimicking these samples to obtain the reliable last cluster. We summarize the aCL method based on the bootstrap method as below.

Algorithm Prediction (one-step ahead) using the aCL

1: Given a set $\mathbf{Z}_N = \{Z_n : Z_n = (\mathbf{x}_n, y_n) \in \mathbb{R}^{m+1}, n = 1, \dots, N\}$

2: **for** $i \in \{1, \dots, m\}$ **do**

3: Reproduce the i-th bootstrap dataset,
$\mathbf{Z}_N^i = \{Z_N^i : Z_N^i = (\mathbf{x}_{in}, y_{in}) \in \mathbb{R}^{m+1}, n = 1, \dots, N\}$

4: Obtain a last cluster by partitioning the dataset (such as k-means clustering), and denote the size of the last cluster as L^i

5: Let the size of aCL be $C^i = \alpha_i L^i + (1 - \alpha_i)N, \ \ 0 \le \alpha_i \le 1$

6: Estimate $\hat{C}^i$ by minimizing the value of the risk function,
$\hat{C}^i = \arg\min_{L^i \le C^i \le N} \frac{1}{C^i} \sum_{n=N-C^i+1}^{N} \left(y_{in} - \hat{f}(\mathbf{x}_{in})\right)^2$

7: **end for**

8: Calculate the average of the $\hat{C}^i$'s, $\bar{C} = \frac{1}{m}\sum_{i=1}^{m} \hat{C}^i$

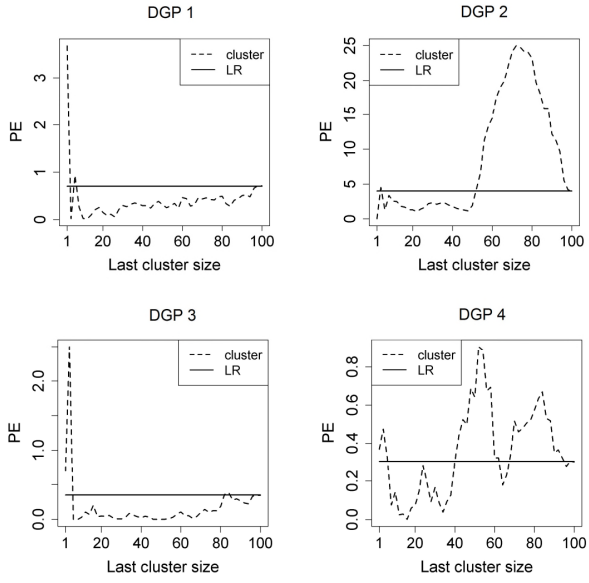
10: Estimate the regression coefficient using $\bar{C}$,
$\hat{\beta}_{\bar{C}} = \arg\min_{\beta_{\bar{C}} \in \mathbb{R}^m} \frac{1}{\bar{C}} \sum_{n=N-\bar{C}+1}^{N} \left(y_n - \hat{f}(\mathbf{x}_n)\right)^2$

11: Using $\hat{\beta}_{\bar{C}}$, construct the predictor $\hat{f}(\mathbf{x}_{N+1})$ for y_{N+1}

[Fig. 3] Algorithm of aCL method

3. Results

To examine whether the aCL method applies well to various functional forms, we conducted Monte Carlo simulations over four types of DGPs. We can observe that prediction based on some last clusters results in lower prediction error than does prediction using the entire dataset. Nevertheless, there exist other last clusters that yield higher prediction error than using the whole dataset. Therefore, selecting a last cluster of adequate size is important and so using the aCL method can be useful for obtaining the last cluster.



[Fig. 4] Change of the prediction error based on the size of the last cluster.

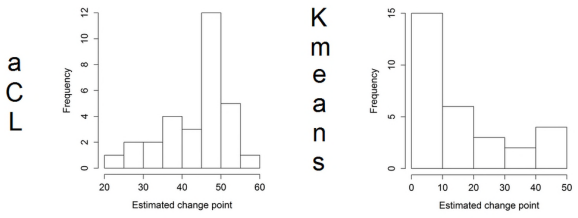
Table 1. shows the prediction errors of the methods used in this study. The results in Table 1. indicate that the three aCL methods outperform other different prediction methods under different DGPs.

[Table 1] Comparison of prediction error

DGP	1			2		
Variance	1	2	3	1	2	3
OLS	245.019	348.978	342.525	1.312	1.262	3.423
OLSaCL	11.608	14.307	17.168	0.313	1.162	3.69
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.4	-0.599
LSPD	246.171	346.157	340.285	2.421	2.595	4.819
LSPDaCL	20.039	27.033	28.1	0.571	1.661	4.098
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.039	-0.302
LASSO	243.205	341.971	339.174	1.287	1.175	3.27
LASaCL	12.787	19.905	18.788	0.424	1.354	4.409
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.406	-0.563
Kmeans	31.493	57.695	56.587	1.108	2.673	7.525
(p-value)	-0.002	(<0.001)	(<0.001)	-0.001	-0.051	-0.014
Bootstr	242.343	346.982	341.665	1.308	1.314	3.334
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.349	-0.638
Comb	176.012	274.632	282.323	0.895	1.201	3.818
(p-value)	(<0.001)	(<0.001)	(<0.001)	-0.002	-0.461	-0.457

DGP	3			4		
Variance	1	2	3	1	2	3
OLS	11.451	14.406	20.64	7.727	7.565	7.016
OLSaCL	1.264	2.792	6.841	0.752	1.609	3.186
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.001
LSPD	11.537	14.523	20.668	7.645	7.669	7.386
LSPDaCL	1.933	3.41	7.432	1.492	2.258	3.612
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.001
LASSO	10.674	13.425	20.546	7.798	8.409	7.049
LASaCL	1.463	2.781	5.935	0.888	1.239	2.863
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.001
Kmeans	2.563	7.893	10.698	1.822	4.847	13.545
(p-value)	-0.004	(<0.001)	-0.018	-0.001	(<0.001)	(<0.001)
Bootstr	11.041	14.159	20.696	7.654	7.491	7.192
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.001
Comb	10.029	12.552	15.673	5.273	5.866	6.986
(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	-0.001

Fig 5. depicts the histograms of estimated change points by aCL and Kmeans through 100 Monte Carlo iterations. We can see that the histogram of aCL shows a central tendency similar to a normal distribution, unlike Kmeans.



[Fig. 5] Histograms of change points estimated by aCL and Kmeans.

We also tested the prediction power of the aCL method using two real-world datasets, stock market index and home price index. The prediction results also showed that the aCL method can improve prediction performance compared to other competing

methods (In this version of paper, the results are omitted for the page limit).

4. Conclusion

Partitioning data produces a more accurate prediction than using a whole dataset when a true function is not known or when setting an estimator for the whole dataset is difficult. Meanwhile, the small-sized cluster produced by simply partitioning data causes high variance problems, which leads to high prediction errors. In this study, we showed that adjusting the size of the last cluster avoids high prediction error. To adjust the size of the last cluster between the whole dataset and the last cluster, we applied the bootstrap method, which has the effect that a given model is trained multiple times by different bootstrapped samples generated from a certain DGP; hence, the model can estimate the reliable location of the last cluster that is optimal for prediction. Therefore, we believe that this paper contributes to understanding that the adjusted last cluster could be optimal for prediction based on the balance between bias and variance by the bootstrap method.

As a result, the aCL method was shown to reduce prediction errors using the numerical results of both simulated and real data. Notice that the aCL method is not for establishing a complete model fitting a whole dataset but for selecting an optimal subsample for prediction. Therefore, an estimator of simple functional form such as a linear function can be easily used with the aCL method for prediction, and this advantage of the aCL method would serve a practical use in research or application. In addition, the aCL method yielded the size-adjusted last cluster by the bootstrap method, which is easy to be adopted in studies using a clustering algorithm, and thus produces a more reliable subsample for prediction.

References

[1] Dietterich TG, Kong EB. "Machine learning bias, statistical bias, and statistical variance of decision tree algorithms", Technical report, Department of Computer Science, Oregon State University, 1995.

[2] Horowitz JL, Handbook of econometrics, p. 3163, 2001.

[3] Breiman L. Bias, variance, and arcing classifiers, Technical report, 460, Statistics Department, University of California, Berkeley, CA, USA, 1996.

[4] Kim JW, Kim JC, Kim JH, "Adjusted k-nearest neighbor algorithm", Journal of the Korean Society of Marine Engineering, Vol. 42, No. 2, pp. 127-135, 2018.

[5] Diebold FX, Chen C. "Testing structural stability with endogenous breakpoint a size comparison of analytic and bootstrap procedures", Journal of Econometrics, Vol. 70, No. 1, pp. 221-241, 1996.