Journal of the Korean Institute of Industrial Engineers
Vol. 42, No. 5, pp. 314-326, October 2016.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2016.42.5.314
© 2016 KIIE
<Original Research Paper>

다중선형회귀모형에서의 변수선택기법 평가

류나현 · 김형석 · 강필성†

고려대학교 산업경영공학부

Evaluating Variable Selection Techniques for Multivariate Linear Regression

Nahyeon Ryu · Hyungseok Kim · Pilsung Kang

School of Industrial Management Engineering, Korea University

The purpose of variable selection techniques is to select a subset of relevant variables for a particular learning algorithm in order to improve the accuracy of prediction model and improve the efficiency of the model. We conduct an empirical analysis to evaluate and compare seven well-known variable selection techniques for multiple linear regression model, which is one of the most commonly used regression model in practice. The variable selection techniques we apply are forward selection, backward elimination, stepwise selection, genetic algorithm (GA), ridge regression, lasso (Least Absolute Shrinkage and Selection Operator) and elastic net. Based on the experiment with 49 regression data sets, it is found that GA resulted in the lowest error rates while lasso most significantly reduces the number of variables. In terms of computational efficiency, forward/backward elimination and lasso requires less time than the other techniques.

Keywords: Forward Selection, Backward Elimination, Stepwise Selection, Genetic Algorithm, Ridge Regression, Lasso, Elastic Net

1. 서 론

다중선형회귀분석(multivariate linear regression : MLR)은 다수의 설명변수(explanatory variables)와 하나의 반응변수(response variable) 사이의 관계를 선형으로 가정하고 주어진 데이터를 가장 잘 설명할 수 있도록 회귀 계수(regression coefficients)를 추정하는 통계적 기법이다(Ross, 2004). 선형회귀분석은 몇 가지 가정을 데이터가 만족할 경우 명시적으로 최적의 해가 존재한다는 점과 더불어 각 설명변수의 통계적 유의성을 도출할 수 있으며 해당 설명변수의 변화가 반응변수에 미치는 영향력을 정량화할 수 있다는 점, 그리고 노이즈에 상대적으로 민감하지 않으면서 과적합의 우려가 적다는 장점 등으로 인하여 여전히 산업 현장에서는 복잡한 최신의 비선형 알고리즘에 비해 실제 시스템을 구현할 때 선호되는 회귀 모형이다(Kang *et al*., 2009; Kang *et al*., 2011; Lastovicka and Sirianni, 2011; Lee *et al*., 2014).

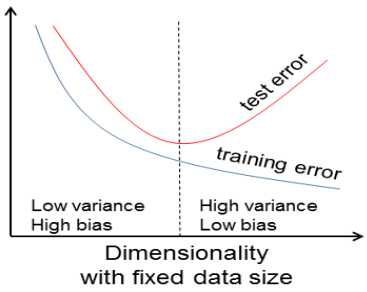

Figure 1. Error rates(y-axis) with the number of variables(x-axis)

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행됨(No. 2014R1A1A1004648, No. NRF-2015R1A2A2A04007359).
† 연락저자 : 강필성 교수, 02841 서울특별시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3383, Fax : 02-929-5888, E-mail : pilsung_kang@korea.ac.kr
2016년 6월 10일 접수; 2016년 7월 22일 수정본 접수; 2016년 10월 4일 게재 확정.

다중선형회귀분석의 가정 중 하나는 설명변수 간 독립성이 확보되어야 한다는 것인데 실제 산업 현장의 데이터에서는 변수가 많아질수록 서로 연관성이 있는 설명변수들이 생길 가능성이 커지기 때문에 설명변수 간 독립성을 확보하는 것이 어려워지는 상황에 자주 직면한다. 또한 설명변수가 증가할수록 회귀모형의 복잡도도 높아지기 때문에 변수가 어느 수준 이상으로 증가하면 회귀모형의 예측 성능이 저하될 가능성이 커지게 된다. <Figure 1>은 데이터에서 관측치가 일정할 때 차원의 증가에 따른 예측 오차(prediction error)의 변화를 설명하고 있다. 저차원에서는 변수가 증가할수록 예측 오차가 감소하는 경향이 있지만 일정 수준 이상의 변수를 사용할 경우 새로운 데이터에 대해서는 오히려 예측 오차가 증가하는 현상을 나타내게 된다. 고차원에서는 차원이 높을수록 일정한 예측 정확성을 위해 필요한 관측치의 수가 급격히 증가하는데 이를 차원의 저주(curse of dimensionality)라고 부른다. 차원이 증가하면 차원의 수에 비해 상대적으로 관측치의 수가 부족하기 때문에 예측 결과의 변동성이 커지므로 차원이 높을수록 고차원에서는 예측 오차는 증가하게 되는 것이다(Bellman, 2015; Madhuri and Rani, 2015). 따라서 제한된 관측치 내에서 차원의 저주를 해결하기 위해서는 예측모델을 구축하기 위한 효율적인 변수 집합을 구성해야 한다.

실제 선형회귀분석에서는 모든 설명변수가 통계적으로 유의미하게 반응변수에 영향을 미치는 것이 아니기 때문에 반응변수에 대한 영향력이 높은 변수를 선택하여 고차원에서 예측 성능이 저하되는 문제를 해소시킬 수 있다(Mallick and Yi, 2013). 변수선택법(variable selection)은 예측 모델에서 유의미한 변수를 선택하고 오차를 야기하거나 불필요한 변수를 제거함으로써 모델의 예측 성능을 향상시킬 수 있으며(Blum and Langley, 1997; Guyon and Elisseeff, 2003) 고차원 데이터에서의 차원축소로 모델의 복잡성을 줄이고 예측의 효율성을 높일 수 있는 방법론이다. 또한 변수선택법은 학습 알고리즘과 관련 분야에 적합된 최적의 변수를 판별할 수 있다는 장점이 있다(Kohavi and John, 1997; Smialowski *et al*., 2010).

지금까지 다중선형회귀분석에 사용될 수 있는 변수선택기법은 다수가 제안되었다. 이들의 이론적인 특징은 상대적으로 잘 알려져 있는 반면에, 현실 문제를 해결함에 있어 각 방법론들을 체계적이고 실증적으로 비교한 실험 연구는 다소 미흡한 실정이다. 따라서 다양한 데이터에 대한 실험을 통하여 여러 변수선택 방법론들을 비교/평가하는 연구는 실용적인 관점에서 큰 의의를 갖는다고 할 수 있다. 이를 위하여 본 연구에서는 다중선형회귀분석에 대해 전진 선택법(forward selection), 후방소거법(backward elimination), 단계적 선택법(stepwise selection), 유전 알고리즘(genetic algorithm), 그리고 축소 추정법인 ridge 회귀(ridge regression), lasso(least absolute shrinkage and selection operator), elastic net, 총 일곱 가지의 변수선택기법들의 결과를 비교 평가하고자 한다. 전체 변수 사용 대비 예측 정확도와 변수 감소 비율, 실행 시간 등의 관점에서 분석을 수행하였으며 이를 위하여 총 49개의 데이터 셋을 사용하였다.

따라서 본 연구를 다음과 같이 구성하였다. 제 2장에서는 변수선택법 학습 알고리즘으로써 다중선형회귀분석에 대하여 설명하고 제 3장에서는 일곱 가지의 변수선택기법들에 대해 기술한다. 제 4장에서는 변수선택법 평가를 위한 실험 방법 및 데이터를 설명하며, 제 5장에서 실험 결과 분석과 변수선택기법 평가에 대해 서술한다. 마지막 제 6장에서는 본 연구의 결론 및 의의를 다룬다.

2. 다중선형회귀분석(Multiple Linear Regression)

회귀분석은 반응변수에 대한 설명변수의 영향력을 예측하는 방법으로 반응변수 변화의 원인을 설명변수로 해석이 가능하다. 설명변수와 반응변수의 관계를 선형적이라 가정한 경우 선형회귀분석이라 하며 이 선형적 관계는 예측 모델을 가장 단순하게 만들고 표본데이터에 의한 과적합을 줄일 수 있다(James *et al*., 2013). 이러한 이유에서인지 실제 산업 현장에서 선형회귀분석이 가장 많이 쓰이고 있으며 실제 데이터는 설명변수 하나만으로 회귀분석을 할 수 있는 경우는 매우 드물기 때문에 다수의 설명변수를 활용한 다중선형회귀분석이 주로 사용된다.

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_\mathrm{p} \mathrm{X}_\mathrm{p} + \epsilon \tag{1}$$

다중선형회귀모형을 일반화하면 위 식과 같으며 여기서 Y는 반응변수, X는 설명변수, β는 선형회귀계수, ε는 오차이다. 이때 ε는 기댓값은 0이고 분산이 일정한 정규분포를 따른다고 가정한다. 선형회귀분석의 회귀 계수를 추정하기 위해서 대표적으로 최소제곱법(ordinary least square; OLS)을 사용하며 다음과 같이 오차 제곱합을 최소화하는 회귀 계수를 추정하게 된다.

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \epsilon_\mathrm{i}^2 \\ &= \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{p} (y - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned} \tag{2}$$

회귀분석은 설명변수와 반응변수의 상관관계가 높고 설명변수가 서로 독립이라는 가정 하에서 이뤄지며 설명변수가 서로 독립적일 때 각 설명변수의 설명력이 높다. 그러나 현실에서 수집된 데이터들은 설명변수들이 완전히 독립인 경우는 매우 드물지만 대부분의 경우 분석에 큰 문제가 되지 않는다. 하지만 설명변수 간 비직교성(nonorthogonality)이 너무 강하면 설명변수의 특성이 서로 유사하여 회귀계수 추정량의 변동이 커지게 된다. 이처럼 설명변수 사이의 연관성이 높아 발생하는 문제를 다중공선성이라 한다(Frisch, 1934). 다중공선성의 발생 원인은 크게 두 가지로 나뉘는데 첫째 관측치가 변수에

비해 적은 경우, 둘째 다중공선성의 발원이 되는 변수 간 상관관계가 고유속성일 경우이다. 전자의 경우 표본을 추가함으로써 오차를 낮추어 문제를 해결할 수 있지만(Chatterjee and Hadi, 2015), 후자의 경우에는 기 존재하는 설명변수로부터 다중공선성이 발생하지 않는 일부의 설명변수 집합을 추출해서 회귀모형을 구축해야 한다.

3. 변수선택법(Variable Selection)

변수선택법은 주어진 모델에서 유의미한 변수를 선택하거나 불필요한 변수를 소거하여 모델에서 사용할 변수의 조합과 차원의 수를 최적화하는 것이 목적으로 일반적인 절차는 <Figure 2>에 나타난 것과 같다. 변수선택의 궁극적인 목적은 예측 모델의 오차를 최소할 수 있는 최소한의 변수 집합을 선택하는 것이다. 본 연구에서는 총 일곱 가지의 변수선택법을 통해 다중선형회귀모델을 학습하고 평가하여 각각의 변수선택법들을 비교 분석하였다. 최적의 변수선택을 위해서는 전역적 탐색(exhaustive search)이 이상적이지만 이 기법은 데이터의 차원이 증가할수록 높은 계산 복잡도가 요구되기 때문에 현실적으로는 사용이 불가능하며 효율적인 지역적 탐색 기법을 통해 변수선택을 최적화할 필요가 있다.

3.1 전진 선택법(Forward Selection)

가장 단순한 변수선택법 중 하나로 설명변수를 하나도 사용하지 않는 모델부터 시작하여 유의미한 설명변수 순으로 설명변수를 하나씩 추가하는 방법으로 변수를 선택하는 기법이다. 전진 선택법에서는 이전 변수 집합에 비해 어느 새로운 변수도 모델 적합도의 증가 폭이 기준치 이상을 나타내지 못할 경우에 변수선택이 종료된다. 전진 선택법은 매우 단순한 모델에서부터 탐색을 시작하기 때문에 변수가 많을 때 빠르게 변수선택을 할 수 있는 장점이 있지만 한번 선택된 변수는 다시 제거되지 않는 Greedy 방식의 탐색 기법이라고 할 수 있다.

3.2 후방 소거법(Backward Elimination)

데이터의 전체 변수를 포함하는 모델에서 유의미하지 않는 순으로 설명변수를 하나씩 제거하는 방법으로 전진 선택법과 반대 방향으로 변수를 선택하는 기법이다. 후방 소거법에서는 기존 변수 집합에서 어느 한 변수를 제거하더라도 모델 적합도가 크게 감소하는 경우에 더 이상 변수를 제거하지 않고 종료가 된다. 전진 선택법과 유사하게 후방 소거법 역시 한번 제거된 변수는 모델링에 다시 고려되지 않는 Greedy 방식의 탐색 기법이며, 변수의 수가 많을 경우에는 학습 초기에 계산 시간이 길어질 수 있어 전진 선택법에 비해 느릴 수 있다. 하지만 필요한 변수를 제거의 대상에서 제외시키면서 변수를 선택하기 때문에 유의미한 변수가 소거될 가능성이 적어져 상대적으로 안전한 방법론이라 할 수 있다.

3.3 단계적 선택법(Stepwise Selection)

앞의 두 방법의 단점을 보완한 방법으로 한 번 선택되거나 소거된 설명 변수를 다시 고려하여 양방향으로 변수를 선택할 수 있는 방법이다. 즉 학습 초기에 전진 선택법과 같은 방법으로 시작하였다가 선택된 변수가 3개 이상이 되면 양방향으로 전진 선택법과 후방 소거법을 번갈아 시행한다. 단계적 선택법은 한번 선택되거나 소거된 변수가 고정되지 않기 때문에 결과가 비교적 안정적이며 이전의 전진선택법과 후방소거법보다 상대적으로 많은 경우의 수를 탐색함으로써 보다 최적의 결과를 얻을 수 있다. 그러나 더 많은 경우의 수를 계산하는 만큼 더 긴 시간이 소요되어 앞의 두 방법론들보다 학습이 느리다는 단점이 있다.

3.4 유전 알고리즘

<Figure 3>은 유전 알고리즘의 일반적인 절차를 나타낸 것이다. 본래 유전 알고리즘은 생물의 진화론과 유전 법칙을 모방하여 최적의 해를 찾는 휴리스틱 방법 중 하나로써 유전 알고리즘의 변수선택은 가능한 조합의 설명변수들을 염색체 형태로 표현하여 최적의 설명변수 조합을 찾도록 한 기법이다(Yang and Honavar, 1998). 유전 알고리즘의 유전자(gene)는 변수선택법에서의 변수를 표현하며 유전자의 값이 1이면 해당 변수를 선택한다는 것을 의미하고 값이 0이면 해당 변수를 모델링에서 제외시킨다는 것을 의미한다. 염색체(chromosome)는 전체 설명변수의 유전자로 이루어져 가능한 설명변수의 조합들을 표현하며 염색체의 길이는 전체 설명변수의 수와 같다. 유전 알고리즘은 시작할 때 주어진 세대 크기만큼의 염색체를 임의

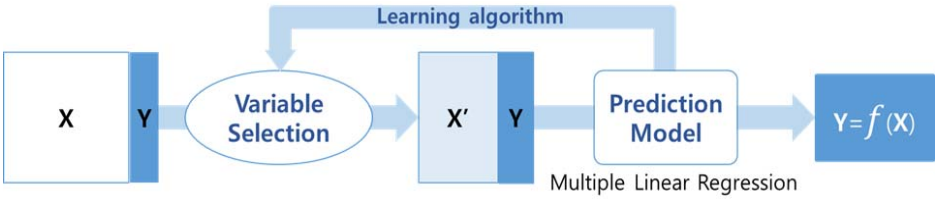

Figure 2. Variable selection procedure

로 생성한다. 생성된 염색체 중 적합도가 높은 염색체를 부모 염색체로 사용하며 부모 염색체는 재생성(reproduction)으로 자손 염색체를 만들어 낸다. 재생성 과정은 <Figure 3>처럼 교차(crossover)와 변이(mutation)로 구분된다. 교차는 부모 염색체로부터 가능한 여러 조합의 설명변수를 만들어 보다 나은 변수 집합을 탐색할 수 있도록 하는 역할을 수행하며, 변이는 부모 염색체로부터 나오는 변수의 조합에 약간의 우연적 요소를 가미한 변화를 시킴으로써 새로운 변수의 조합이 기존 변수 집합과는 다른 영역을 탐색할 수 있는 가능성을 부여하는 역할을 수행한다. <Figure 3>(b)는 가까운 최적 해를 찾는 것을 표현한 것으로 유전 알고리즘에서 교차 재생성에 해당하며 (c)는 지역적 해로부터 빠져나와 전역적 해에 도달하는 변이의 역할을 개념적으로 설명하는 것이다. 세대 크기만큼의 자손 염색체가 재생성 되면 한 세대를 마치고 재생성된 염색체는 다시 적합도 함수에 의하여 다음 세대의 부모 염색체가 된다. 적합도 함수(fitness function)는 염색체의 적합도를 평가하여 한 세대의 염색체 중 모델에 가장 적합한 염색체를 찾도록 하는 함수로 다른 방법론들에서의 목적식과 유사하다. 변수선택을 할 때에는 오차가 가장 작은 경우의 설명변수를 선택하도록 하며 본 실험에서는 다른 기법과 마찬가지로 예측 모델의 성능 지표가 되는 BIC를 최소화 하는 변수 집합을 찾도록 적합도 함수를 설정하였다. 이러한 과정은 미리 정의된 수만큼의 세대를 거쳐 반복되고 세대를 모두 마치면 유전 알고리즘은 완료된다. 그러므로 유전 알고리즘에서는 적합도가 높은 부모 염색체의 유전자(설명변수)가 자손 염색체에 계속 유전되며 마지막 자손 염색체는 모델에 최적인 설명변수들을 나타낸다.

- 축소추정 기법(Shrinkage Method)

전진 선택법, 후방 소거법, 단계적 선택법과 같은 방식의 회귀계수를 추정하는 방식은 회귀계수를 고정한 상태로 여러 변수 집합을 고려하는 기법이다. 이와는 다른 방식의 접근법으로써 축소추정 기법은 회귀식의 목적함수에 회귀 계수에 대한 제약(regularization)을 추가하는 방식으로 선형회귀계수의 크기를 감소시키는 방식이다. 이렇게 추정된 회귀식의 경우 유의미하지 않은 변수는 0의 값을 갖거나 0에 가까운 매우 작은 값을 갖게 되어 변수를 선택할 수 있는 기준을 제공해줄 수 있다. 주로 사용되는 축소 추정법으로는 ridge 회귀, lasso, elastic net 등이 있다.

3.5 Ridge 회귀(Ridge Regression)

다중공선성 문제가 발생하는 고차원 데이터에서는 예측 결과가 불안정하고 분산이 높기 때문에 오차가 크게 나타나는 문제가 발생한다. 이를 해결하기 위해 Hoerl and Kennard(1970)는 다중공선성 문제에서 선형회귀계수의 크기를 감소시켜 모델의 편의(bias)를 허용함으로써 예측 정밀도(prediction precision)를 향상시킬 수 있는 ridge 회귀를 제안하였다. 다음 식은 ridge 회귀의 목적식이며 <Figure 4>에서 alpha가 0인 경우로 나타낼 수 있다.

$$\hat{\beta} = \mathrm{argmin}_{\beta} |\mathbf{Y} - \mathbf{X}\beta|^2 + \lambda_1 |\beta|^2 \tag{3}$$

ridge 회귀의 목적식은 다중선형회귀 분석의 목적식에 L2-norm의 제약 조건이 있는 것으로 다중선형회귀 모형의 선형 계수 값의 크기를 감소시켜 다중공선성에 의한 예측 오차를 줄이게 된다. 이때 선형 계수 값의 크기는 무한히 작아질 수 있지만 0이 될 수는 없다. 따라서 본 연구에서는 일정 임계치 이하의 작은 계수를 가지는 변수들을 소거를 통한 ridge 기반의 변수선택법을 사용하였다.

3.6 Lasso(Least Absolute Shrinkage and Selection Operator)

ridge 회귀의 선형 계수는 0이 될 수 없으므로 변수선택이 불

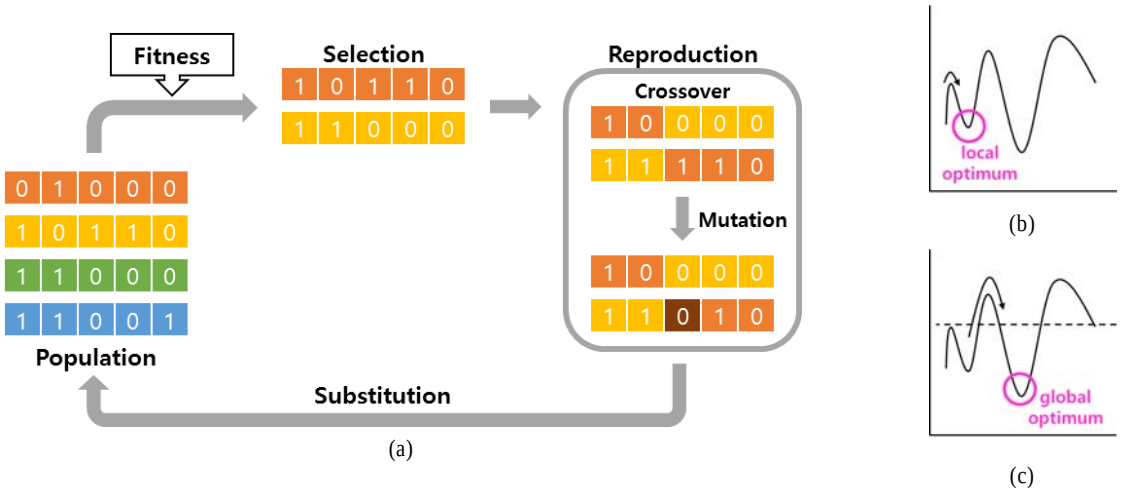

Figure 3. Genetic algorithm process for variable selection

가능하다는 단점이 있어 Tibshirani(1996)에 의하여 lasso가 제안되었다. lasso는 ridge 회귀를 보완하여 변수선택이 가능하도록 한 축소추정 방법이며 다음 식을 목적식으로 한다. <Figure 4>에서는 alpha가 1인 경우에 해당한다.

$$\hat{\beta} = \mathrm{argmin}_{\beta} |\mathbf{Y} - \mathbf{X}\beta|^2 + \lambda_2 |\beta|^1 \quad (4)$$

Lasso 목적식에서 제약 조건인 L1-norm regularizer는 선형 계수 값의 크기를 감소시키거나 0이 될 수 있도록 교정한다. 그래서 lasso는 ridge 회귀와 차원 축소의 장점을 모두 채택한 방법으로 고안되었다.

<Figure 4>는 회귀계수의 좌표 상에서 ridge 회귀, lasso, elastic net의 제약식과 회귀계수 추정치 $\hat{\beta}$의 분포를 그린 것이다. Ridge 회귀의 L2-norm regularizer는 구(球) 모양으로 원점으로부터 항상 동일한 제약 조건을 나타낸다. 때문에 ridge 회귀와 회귀계수 분포의 접점은 절편이 아닌 구 표면의 한 점이 될 가능성이 매우 높다. 하지만 lasso의 L1-norm regularizer는 원점과의 거리가 절편에서 가장 멀기 때문에 회귀계수 분포와의 접점이 주로 접점이 된다. 따라서 lasso의 회귀계수는 0이 되는 경향이 크다.

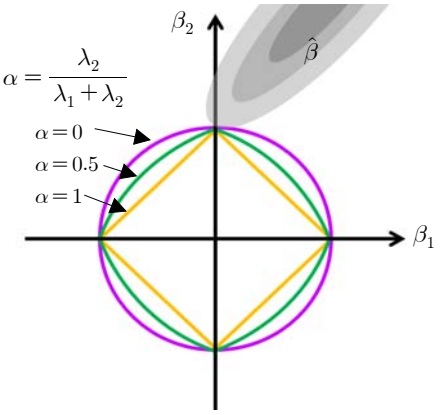

Figure 4. Comparison of different penalties of shrinkage methods

3.7 Elastic Net

설명변수가 매우 많은데 반해 관측치의 수가 매우 적을 경우 설명변수들 사이에 높은 상관관계가 생길 수 있으며 높은 상관관계가 있는 설명변수들은 각각 그룹을 이룰 수 있다. 그런데 만약 그룹 내에 있는 일부 변수를 제외하고 예측을 한다면 그룹 내 제외되지 않은 다른 변수들의 정보가 많이 손실될 우려가 있으며 유의미한 설명변수의 영향이 저하될 수 있다. elastic net은 lasso를 보완하여 긴밀한 설명변수의 그룹을 유지하면서 변수를 선택할 수 있는 방법론을 일컫는다. elastic net의 목적식은 다음과 같다.

$$\hat{\beta} = \mathrm{argmin}_{\beta} |\mathbf{Y} - \mathbf{X}\beta|^2 + \lambda_2 |\beta|^1 + \lambda_1 |\beta|^2 \quad (5)$$

이 목적식은 L1-norm과 L2-norm의 합인 제약 조건을 가지며 이 제약식은 선형 계수 값의 크기를 줄이고 그룹이 유지되도록 변수를 선택(grouped selection)하게 만든다. Elastic net은 <Figure 4>에서 alpha가 0에서 1사이의 수이며 alpha가 1에 가까울수록 lasso와 유사하며 0에 가까울수록 ridge와 유사하다. 본 연구에서는 alpha가 0.5인 elastic net을 이용하여 실험을 수행하였다.

3. 실험 설계

본 연구에서 수행한 다중선형회귀분석에서의 변수선택기법의 효과 검증 절차는 <Figure 5>에 나타난 바와 같다. 데이터는 <Table 1>에 나타난 바와 같이 UCI(http://archive.ics.uci.edu/ml/) 데이터 저장소에서 22개, KEEL(http://sci2s.ugr.es/keel/datasets.php) 데이터 저장소에서 21개, 그리고 R(http://www.r-project.org/) 및 기타 원천으로부터 7개 총 49개의 다중선형회귀분석이 가능한 데이터 셋을 수집하여 실험에 사용하였다. 설명변수 개수는 최소 7개, 최대 95개이고 관측치 개수는 최소 32개, 최대 515,344개이다. 각 데이터 셋에서 무작위로 전체 관측치의 2/3씩 추출하고 표본을 구성하여 학습을 수행하고 나머지 1/3을 검증 데이터로 사용하여 예측 오차율을 산출하였다. 통계적 유의성을 확보하기 위하여 이 절차를 각 데이터셋-변수선택기법마다 30회 반복 실험을 수행하였으며, 유전 알고리즘에 대해서만 10회 반복 실험을 수행하였다. 실험 및 전처리에는 R을 사용하였고 유전 알고리즘의 실험 설정 조건으로는 세대 반복 수 10번, 세대 크기 50개, 교차 재생성 비율 0.8, 변이 비율 0.1을 사용하였다. 축소 추정 기법의 설정 조건으로는 총 10,000개의 λ 중 각 데이터셋 별로 최적의 값을 선택하여 사용

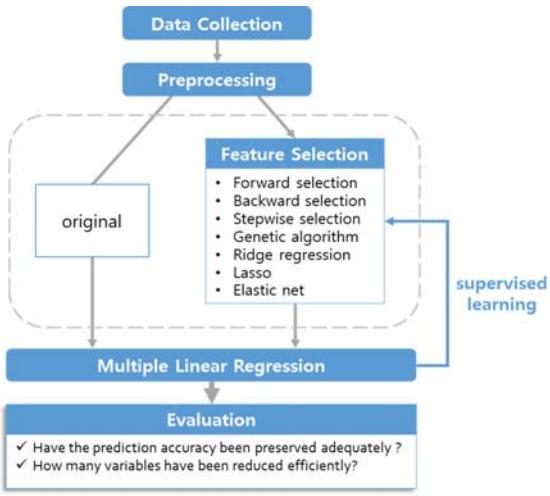

Figure 5. Process of evaluating variable selection techniques for multivariate linear regression

Table 1. Description of the datasets

Dataset	source	records	variables	Dataset	source	records	variables
abalone	KEEL	4,177	9	OVCTT_bitrate	UCI	68,784	16
akbilgic	UCI	536	8	OVCTT_framerate	UCI	68,784	16
ANACALT	KEEL	4,052	8	OVCTT_height	UCI	68,784	16
autoMPG8	KEEL	392	8	OVCTT_width	UCI	68,784	16
baseball	KEEL	336	17	ParkinsonsTele	UCI	5,875	21
Buzz	UCI	28,179	95	PersonalLoan	etc.	2,500	13
california	KEEL	20,640	9	PCASP	UCI	45,730	10
Carseats	R	400	11	pole	KEEL	14,998	27
CBM_C	UCI	11,934	15	puma32h	KEEL	4,124	33
CBM_T	UCI	11,934	15	SML2010	UCI	4,137	24
compactiv	KEEL	8,192	22	solar flare_C	UCI	323	11
concrete	KEEL	1,030	9	solar flare_M	UCI	323	11
dee	KEEL	365	7	solar flare_X	UCI	323	11
delta_elv	KEEL	9,517	7	stock	KEEL	950	10
elevators	KEEL	16,599	19	student	UCI	382	51
EnergyEfficiency_C	UCI	768	9	tic	KEEL	9,822	86
EnergyEfficiency_H	UCI	768	9	ToyotaCorolla	etc.	1,436	34
forestFires	KEEL	517	13	UScrime	R	47	16
Hitters	R	263	20	Wage	R	3,000	10
house	KEEL	22,784	17	wankara	KEEL	1,609	10
housing	UCI	506	14	winequality	UCI	6,497	12
machineCPU	KEEL	209	7	wizmir	KEEL	1,461	10
mortgage	KEEL	1,049	16	wpbc	UCI	194	34
mv	KEEL	40,768	11	yacht	UCI	308	7
nuclear	R	32	11				

하였다. λ는 축소추정을 할 때 회귀계수를 얼마나 축소할지 설정하는데 쓰이는 값으로 λ가 클수록 회귀계수 크기가 많이 줄어들도록 되어 있다. 그 외 다른 설정 값과 전진 선택법, 후방소거법, 단계적 선택법의 설정 값은 R의 기본 설정 값을 사용하였으며 모든 변수선택기법에 대하여 Bayesian Information Criterion(BIC)를 변수선택 기준으로 설정하였다.

변수선택법은 예측 모델의 오차가 최소인 경우 변수를 선택하는데 일반적으로 모델의 학습 오류(training error)는 새로운 데이터에 대한 검증 오류(test error)보다 낮은 경향을 보인다. 따라서 학습 오류에 제약을 더하여 이를 보완한 지표들이 변수선택의 기준으로써 주로 사용된다(James *et al*., 2013). 일반적으로 잘 알려진 변수선택의 기준으로는 Akaike information criterion(AIC), BIC, 수정 R-제곱합(adjusted R2)가 있다. BIC는 데이터의 차원이 높을수록 AIC보다 더 적은 변수를 선택할 수 있도록 한 지표이고(Shumway *et al*., 2010), adjusted R2는 유의하지 않은 변수가 상대적으로 적을 때 사용하는 지표이다(James *et al*., 2013). 따라서 BIC는 모델이 복잡할 때 더 단순한 모델로 차원을 축소시킬 수 있고 유의하지 않은 변수가 적지 않을 때 모델의 예측 성능을 더 향상시킬 수 있기 때문에 본 실험에서는 BIC를 기준으로 변수를 선택하였다. 즉 각 변수선택기법의 변수선택 단계에서 BIC의 값이 최소가 될 때의 변수를 선택하여 그 차이를 비교 분석하였다. BIC는 다음 식 (6)과 같다.

$$\mathrm{BIC} = \frac{1}{\mathrm{n}}(\mathrm{RSS} + d\hat{\sigma}^2 \ln(n)) \tag{6}$$

식 (6)에서 n은 관측치의 수이며 d는 변수의 수, $\hat{\sigma^2}$은 표본분산, RSS(Residual Sum of Squares)는 오차의 제곱합으로 $\sum_{i=1}^{n} \epsilon_i^2$을 의미한다.

변수선택기법의 평가 기준은 오류 개선율(error rate improvement)과 변수 감소 비율(variable reduction)로 각각 예측 정확도 측면과 모델의 효율성 측면에 대한 지표이며 다음 식 (7)과 식 (8) 같이 정의하였다.

$$\text{error rate improvement} = -\frac{error_{after} - error_{before}}{error_{before}} \times 100 \quad (7)$$

$$\text{variable reduction} = \frac{\text{original dimensionality} - \text{reduced dim}ensionality}{\text{original dim}ensionality} \quad (8)$$

4. 변수선택법 평가

<Figure 6>은 모델의 예측력 측면에서 각 모델의 오류 개선율을 나타낸 것으로, 모든 변수를 사용한 회귀모형에 비해 오류가 감소할수록 진한 파란 색을 나타내고 오류가 증가할수록 진한 붉은색으로, 모든 변수를 사용한 회귀 모형과 변수선택을 수행한 회귀 모형 사이의 차이가 적을수록 흰색에 가깝게 나타낸 것이다. 원래 모델의 절대 오류율이 매우 작아서 변수선택 후 오류율이 두 배 이상(100% 이상) 증가한 경우에는 -100의 값을 대신 사용하였다. <Figure 6>의 결과를 바탕으로 다음과 같은 실험적 결론을 내릴 수 있다. 첫째, 축소 추정법은 변수선택을 수행할 경우 다른 변수선택 기법에 비해 회귀모형의 성능이 다소 저하되는 경향을 보인다. 전진 선택법이나 후방 소거법의 경우 거의 모든 데이터에 대해서 모든 변수를 사용한 회귀모형과 예측력 측면에서 차이를 나타내지 않으나, 축소추정법, 특히 lasso와 elastic net의 경우 상당히 많은 데이터셋에서 상대적으로 낮은 예측 성능을 나타내고 있다. 둘째, 변수선택을 위한 탐색 영역이 증가할수록 회귀모형의 성능이 향상될 가능성이 증가한다. 단계적 선택법이나 유전 알고리즘의 경우 전진 선택법과 후방 소거법 대비 예측력이 향상된 데이터셋의

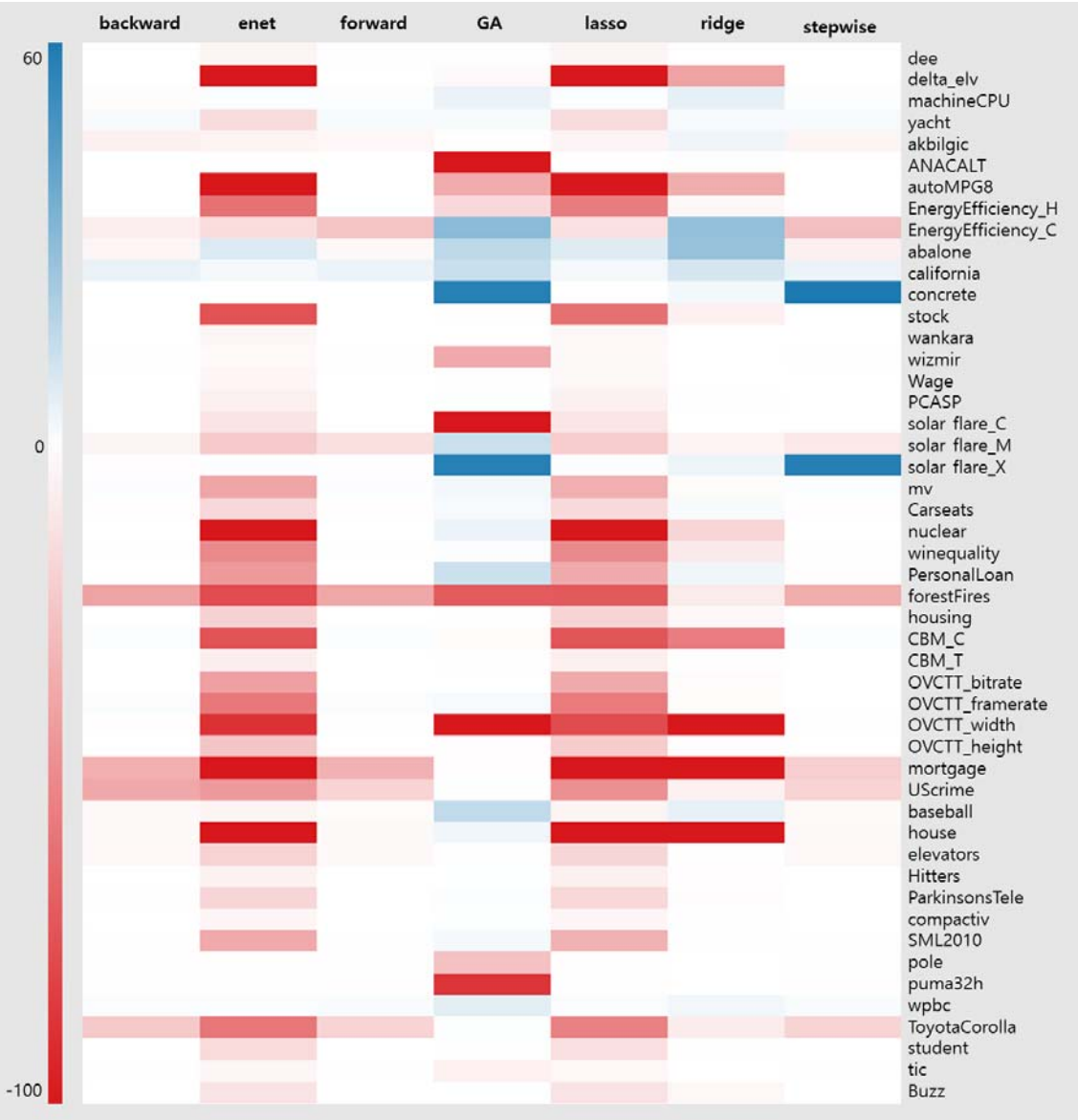

Figure 6. Error rate improvement of variable selection technique for each dataset

수가 상대적으로 많은 것을 알 수 있다. 다만 유전 알고리즘의 경우 예측력의 편차가 단계적 선택법에 비해 크게 나타나는데, 이는 실험 환경에서 세대의 반복수를 상대적으로 적게 제한한 것이 한 가지 원인으로 지목될 수 있다.

<Table 2>는 각 변수선택기법 사이의 오류 개선율에 대한 통계적 유의성을 윌콕슨 순위합 검정을 이용하여 수행한 결과이다. 각 셀의 변수선택기법은 두 변수선택기법 중 오류 개선율이 우수한 변수선택기법을 의미하며, 숫자는 검정의 유의확률(p-value)을 의미한다. <Table 2>의 검증 결과를 볼 때, 평균적으로 GA의 오류 개선율이 가장 높게 나타나나, 여타 기법과의 통계적인 차이는 유의미하지 않은 것으로 나타났다. 반면에 전진 선택법, 후방 소거법, 단계적 선택법의 경우 lasso와 elastic net에 비해 통계적으로 유의미하게 오류 개선율이 높다는 것을 확인할 수 있다. 이 세 가지 변수선택기법 사이에서는 단계적 선택법-후방 소거법-전진 선택법 순으로 오류 개선율이 높게 나타났으나, 이 차이에 대한 충분한 통계적 유의성은 확보되지 않는 것으로 나타났다.

<Figure 7>은 변수선택의 효율성 측면에서 변수 감소 비율을 나타낸 것으로 변수 감소 비율이 높을수록 진한 녹색이 되며, 변수 감소 비율이 낮을수록 흰색에 가깝도록 나타내었다. 변수선택의 효율성 측면에서 판단할 경우 다음과 같은 결론을 내릴 수 있다. 첫째, 축소추정 기반 변수선택법 중에서는 lasso가 가장 효율적으로 변수의 수를 감소시키며, ridge 회귀의 경우 상대적으로 변수 감소 비율이 낮은 것을 확인할 수 있다. lasso의 경우 회귀식의 목적함수에 L1-norm을 적용함으로써 상대적으로 회귀 계수의 값이 0이 되는 상황을 선호하게 되는데 본 연구의 실험 결과는 그 목적이 정확히 달성된 것으로 판단할 수 있다. 반면 ridge 회귀는 대부분의 데이터 셋에서 거의 모든 변수를 그대로 사용하는 것을 알 수 있는데, 이는 ridge 회귀 자체가 회귀 계수를 0으로 만드는 것을 목적으로 하는 것이 아니기 때문으로 유추해볼 수 있다. 둘째, 비축소추정 기반 변수선택법 중에서는 유전 알고리즘의 변수 감소 비율이 가장 낮게 나타나는 것을 알 수 있다. 이는 유전자 알고리즘을 이용하여 변수선택을 수행한 다수의 실제 사례에서도 확인할 수 있는 현상으로써(Kang *et al*., 2009; Kang *et al*., 2011; Kim *et al*., 2015), 탐색 영역이 확장되면서 보다 많은 변수가 선택되는 것으로 유추해볼 수 있다. 마지막으로 비축소추정 기반 변수선택법 중에서 가장 효율적으로 변수의 수를 감소시킬 수 있는 방법론은 단계적 선택법인 것으로 나타났다. 단계적 선택법은 일반적으로 전진 선택법에 의한 변수의 수보다 같거나 적은 수의 변수를 선택하기 때문에 세 가지 변수선택기법 중 가장 효율적으로 변수선택을 수행하는 것으로 확인되었다.

<Table 3>은 각 변수선택기법 사이의 변수 감소율에 대한 통계적 유의성을 윌콕슨 순위합 검정을 이용하여 수행한 결과이다. 각 셀의 변수선택기법 및 숫자의 의미는 <Table 2>와 동일하다. <Table 3>의 검증결과를 볼 때 lasso가 모든 변수선택기법 중에서 가장 효율적으로 변수선택을 수행한 것을 알 수 있다. 또한 GA를 제외한 비축소추정 기반 변수선택기법인 단계적 선택법, 전진 선택법, 후방 소거법은 GA와 ridge 회귀, elastic net에 비해 변수 감소 비율이 통계적으로 유의미하게 높게 나타났다. 이 세 기법 사이에서는 단계적 선택법-후방소거법-전진선택법 순으로 변수 감소 비율이 높은 것으로 나타났으나 통계적으로 유의미한 차이를 나타낸다고는 볼 수 없다.

<Table 4>에서는 앞서 살펴본 회귀모형의 예측 성능 및 변수 감소 비율과 함께 계산 시간 효율성을 고려하여 각 기법간의 상대적 순위를 선정한 것이다. 계산 시간 효율성은 각 변수선택기법이 실험을 통해 변수선택을 하는데 걸리는 평균 시간을 비교한 결과이다. 예측 정확도 향상의 측면에서는 유전자 알고리즘-ridge 회귀-단계적 선택법 순으로 효과적인 것으로 나타났으며, 변수 감소율 측면에서는 lasso-단계적 선택법-후방 소거법 순으로 효율적인 것으로 나타났다. 또한 계산 시간 효율성 관점에서는 전진 선택법-후방 소거법-lasso의 순으로 빠

Table 2. Rank-sum test of error rate improvement

	Forward	Backward	Stepwise	GA	Ridge	Lasso	Elastic Net
Forward		Backward 0.382	Stepwise 0.644	GA 0.974	Ridge 0.246	**Forward 2.16e-8**	**Forward 1.99e-8**
Backward	Backward 0.382		Stepwise 0.738	GA 0.983	Ridge 0.294	**backward 3.70e-8**	**Backward 2.33e-8**
Stepwise	Stepwise 0.644	Stepwise 0.738		GA 0.935	Ridge 0.168	**stepwise 2.13e-8**	**Stepwise 1.82e-8**
GA	GA 0.974	GA 0.983	GA 0.935		GA 1.000	GA 0.962	GA 1.000
Ridge	Ridge 0.246	Ridge 0.294	Ridge 0.168	GA 1.000		Ridge 1.000	Ridge 2.10e-6
Lasso	**Forward 2.16e-8**	**Backward 3.70e-8**	**Stepwise 2.13e-8**	GA 0.962	Ridge 1.000		ENet 0.398
Elastic Net	**Forward 1.99e-8**	**Backward 2.33e-8**	**Stepwise 1.82e-8**	GA 1.000	Ridge 2.10e-6	ENet 0.398	

Figure 7. Variable reduction rate of variable selection technique for each dataset

Table 3. Rank-sum test of variable reduction rate

	Forward	Backward	Stepwise	GA	Ridge	Lasso	Elastic Net
Forward		Backward 0.481	Stepwise 0.470	**Forward 0.000**	**Forward 6.45e-9**	Lasso 0.251	Forward 0.026
Backward	Backward 0.481		Stepwise 0.486	**Backward 0.000**	**Backward 5.04e-9**	Lasso 0.286	Backward 0.032
Stepwise	Stepwise 0.470	Stepwise 0.486		**Stepwise 0.001**	**Stepwise 1.95e-8**	Lasso 0.272	Stepwise 0.034
GA	**Forward 0.000**	**Backward 0.000**	**Stepwise 0.001**		GA 1.000	Lasso 0.000	ENet 0.027
Ridge	**Forward 6.45e-9**	**Backward 5.04e-9**	**Stepwise 1.95e-8**	GA 1.000		Lasso 9.35e-10	ENet 1.000
Lasso	Lasso 0.251	Lasso 0.286	Lasso 0.272	Lasso 0.000	Lasso 9.35e-10		Lasso 0.043
Elastic Net	**Forward 0.026**	**Backward 0.032**	**Stepwise 0.034**	ENet 0.027	ENet 1.000	Lasso 0.043	

Table 4. Rankings of each variable selection techniques for three criteria

Variable selection technique	Error rate improvement	Variable reduction rate	Computational efficiency
Forward	5	4	1
Backward	4	3	2
Stepwise	3	2	6
GA	1	6	7
Ridge	2	7	5
Lasso	7	1	3
Enet	6	5	4

르게 결과를 얻을 수 있는 것으로 확인되었다.

베이지안 이론에서 새로운 정보의 유입은 사전 정보(prior)에 변화를 주어 사후확률(posterior)에 따른 의사결정에 영향을 미친다. 의사결정 시 새로운 정보가 유입되면 사실에 대한 불확실성이 줄어들어 최적에 가까운 의사 결정을 내리는 것을 가능하게 한다. 본 연구는 최적의 변수선택 기법 선택 시 내려지는 결정에 대한 불확실성을 줄여줄 것으로 사료된다. 예를 들어, 본 연구에서 GA는 최소 약 2,266초(약 38분)에서 최대 약 34,421초(약 10시간)의 시간이 변수선택을 하는데 소요되었다. 세대 반복수를 10번으로 제한하여 걸린 시간으로 실제 연구자가 사용할 데이터에 최적화할 만큼 충분히 세대를 반복한다면 더 많은 시간이 소요될 것이다. 만일 연구자가 사용할 데이터에 적합한 기법을 찾기 위하여 모든 변수선택 기법을 적용해 본다면 더욱더 많은 시간이 소요될 것이다. 연구자가 계산 시간 효율성을 고려한다면 예측 정확도 측면에서 ridge 회귀(최소 약 0.04초, 최대 약 1.92초)와 단계적 선택법(최소 약 0.01초, 최대 약 5시간)의 적용을 본 연구를 통해 우선적으로 고려해 볼 수 있을 것이다. 따라서 본 연구를 바탕으로 과업과 데이터의 속성에 따라 예측 정확도, 변수 감소율, 계산 시간 등의 기준에 대한 우선순위가 확립될 경우, 변수선택 기법을 선택함에 있어 최악의 결정을 제외시키고 현실적으로 보다 최적의 변수선택기법 대안 선택을 가능하게 할 것으로 기대한다.

5. 결 론

본 연구에서는 실제 산업 현장에서 가장 많이 사용되고 있는 다중선형회귀분석 기법에 대하여 여러 가지 변수선택기법을 적용했을 때의 결과를 비교하고자 각 기법의 성능을 예측 정확도 측면과 변수선택의 효율성 측면을 중심으로 실제 49개의 데이터에 대해 총 일곱 가지의 변수선택기법들을 실험적으로 평가하였다.

예측 정확도 측면에서 유전자 알고리즘, ridge 회귀, 단계적 선택법, 후방 소거법, 전진 선택법, elastic net, lasso 순으로 예측력이 우수하게 나타났으며, 변수 감소율 측면에서 lasso, 단계적 선택법, 후방 소거법, 전진 선택법, elastic net, 유전자 알고리즘, ridge 회귀 순으로 변수가 효율적으로 감소됨을 알 수 있었다. 결론적으로 전진 선택법, 후방 소거법, 단계적 선택법 방식이 예측 정확도와 변수 감소율 측면에서 유사한 결과를 도출해 내는 것으로 나타났으며, 고차원 데이터에서의 다중공선성 문제 해결을 위한 ridge 회귀, lasso, elastic net은 비축소추정 기반의 변수선택기법보다 항상 좋은 성능을 나타내지는 않는 것으로 확인되었다. ridge 회귀는 가장 안전하게 변수를 유지하는 반면 높은 정확도를 보였으며, lasso는 예측력이 가장 낮지만 가장 높은 변수 감소율을 보여주어 고려된 변수선택기법의 양 극단에 위치하는 것을 알 수 있었다.

본 연구의 의의는 충분히 많은 수의 데이터셋을 이용하여 다양한 변수선택기법의 성능을 실증적으로 비교・평가했다는 것에 있다. 대부분의 데이터마이닝 및 머신러닝 알고리즘의 계산 복잡도는 모델 자체의 복잡도와 데이터의 복잡도(데이터의 크기)에 복합적으로 영향을 받는다. 최근의 빅데이터 환경에서는 아무리 계산복잡도가 낮은 알고리즘이라고 할지라도 데이터의 크기가 기하급수적으로 증가하는 환경이기 때문에 현실적으로 주어진 시간 내에서 적합한 변수를 선택하는 데 있어 가능한 모든 방법론을 적용해보는 것은 현실적으로 불가능한 상황이다. 이러한 조건 하에서 모델 효율성의 관점(적은 변수를 사용한 모델을 선호)과 예측 성능의 관점(우수한 예측 정확도를 나타내는 모델을 선호)에서 적합한 변수 선택 기법의 우선순위를 부여하는데 본 연구의 결과물이 기초 자료로 사용될 수 있을 것으로 기대한다.

참고문헌

Bellman, R. E. (2015), *Adaptive Control Processes : A Guided Tour*, Princeton university press.

Blum, A. L. and Langley, P. (1997), Selection of relevant features and examples in machine learning, *Artificial Intelligence*, **97**(1), 245-271.

Chatterjee, S. and Hadi, A. S. (2015), R*egression Analysis by Example*, John Wiley and Sons.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014), Do we need hundreds of classifiers to solve real world classification problems, *J. Mach. Learn. Res*, **15**(1), 3133-3181.

Guyon, I. and Elisseeff, A. (2003), An introduction to variable and feature selection, *The Journal of Machine Learning Research*, **3**, 1157-1182.

Hoerl, A. E. and Kennard, R. W. (1970), Ridge regression : Biased estimation

for non orthogonal problems, *Technometrics*, **12**(1), 55-67.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, New York : springer, **112**.

Kang, P., Lee, H., Cho, S., Kim, D., Park, J., and Park, C.-K. (2009), A virtual metrology system for semiconductor manufacturing, *Expert Systems with Applications*, **36**(11), 12554-12561.

Kang, P., Kim, D., Lee, H., Doh, S., and Cho, S. (2011), Virtual metrology for run-to-run control in semiconductor manufacturing, *Expert Systems with Applications*, **38**(3), 2508-2522.

Kim, D., Kang, P., Lee, S.-K., Kang, S., Doh, S., and Cho, S. (2015), Improvement of virtual metrology performance by removing metrology noises in a training dataset, *Pattern Analysis and Applications*, **18**(1), 173-189.

Kohavi, R. and John, G. H. (1997), Wrappers for feature subset selection, *Artificial intelligence*, **97**(1), 273-324.

Lastovicka, J. L. and Sirianni, N. J. (2011), Truly, madly, deeply : Consumers in the throes of material possession love, *Journal of Consumer Research*, **38**(2), 323-342.

Lee, H., Kim, S. G., Park, H.-W., and Kang, P. (2014), Pre-launch new product demand forecasting using the Bass model : A statistical and machine learning-based approach, *Technological Forecasting and Social Change*, **86**, 49-64.

Madhuri, V. H. and Rani, T. S. (2015), Ranking and dimensionality reduction using biclustering, *In Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO)*, 209-226.

Mallick, H. and Yi, N. (2013), Bayesian methods for high dimensional linear models, *Journal of Biometrics and Biostatistics*, **1**(5).

Ross, S. M. (2004), *Introduction to Probability and Statistic for Engineers and Scientists*, Academic Press.

Shumway, R. H. and Stoffer, D. S. (2010), Time series analysis and its applications : with R examples, Springer Science and Business Media.

Smialowski, P., Frishman, D., and Kramer, S. (2010), Pitfalls of supervised feature selection, *Bioinformatics*, **26**(3), 440-443.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, 267- 288.

Yang, J. and Honavar, V. (1998), Feature subset selection using a genetic algorithm, *IEEE Intelligent Systems and Their Applications*, **13**(2), 44-49.

<Appendix> 각 데이터셋에 대한 개별 변수선택 방법론의 예측 오차 및 변수선택 비율

Table A-1. The average and standard deviation of prediction errors of each variable selection techniques for each dataset(part 1)

Data set	Original		Forward		Backward		Stepwise		GA		Ridge		Lasso		Elastic Net	
	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.
abalone	0.676	0.025	0.676	0.025	0.676	0.025	0.676	0.025	0.677	0.018	0.700	0.027	0.757	0.082	0.757	0.084
akbilgic	0.422	0.036	0.417	0.038	0.417	0.038	0.417	0.038	0.429	0.040	0.659	0.036	0.731	0.102	0.734	0.103
ANACALT	0.753	0.026	0.754	0.026	0.754	0.026	0.754	0.026	0.751	0.025	0.755	0.028	0.805	0.071	0.807	0.073
autoMPG8	0.425	0.042	0.422	0.045	0.422	0.045	0.422	0.045	0.413	0.050	0.432	0.049	0.573	0.167	0.589	0.181
baseball	0.590	0.058	0.596	0.054	0.596	0.056	0.595	0.056	0.578	0.066	0.579	0.055	0.684	0.127	0.690	0.130
Buzz	0.196	0.016	0.196	0.016	0.196	0.016	0.196	0.016	0.187	0.018	0.231	0.022	0.426	0.211	0.434	0.215
california	0.603	0.010	0.603	0.010	0.603	0.010	0.603	0.010	0.602	0.014	0.625	0.009	0.717	0.098	0.721	0.105
Carseats	0.369	0.025	0.365	0.024	0.365	0.024	0.369	0.024	0.361	0.031	0.373	0.025	0.580	0.207	0.585	0.206
CBM_C	0.371	0.005	0.373	0.006	0.374	0.006	0.373	0.006	65.430	0.024	0.765	0.013	0.658	0.224	0.701	0.216
CBM_T	0.340	0.005	0.454	0.023	0.457	0.006	0.411	0.060	0.343	0.005	0.916	0.012	0.882	0.137	0.855	0.146
compactiv	0.515	0.021	0.515	0.021	0.515	0.021	0.515	0.021	0.519	0.024	0.520	0.023	0.629	0.128	0.641	0.137
concrete	0.629	0.024	0.629	0.024	0.631	0.024	0.629	0.025	0.623	0.027	0.636	0.023	0.736	0.114	0.740	0.117
dee	0.421	0.041	0.422	0.041	0.422	0.041	0.420	0.040	0.411	0.052	0.425	0.040	0.561	0.162	0.576	0.173
delta_elv	0.606	0.013	0.606	0.013	0.606	0.013	0.606	0.013	0.610	0.017	0.609	0.013	0.689	0.097	0.701	0.104
elevators	0.431	0.013	0.432	0.013	0.432	0.013	0.432	0.013	0.425	0.009	0.470	0.015	0.650	0.175	0.649	0.175
EnergyEfficiency_C	0.335	0.016	0.398	0.017	0.413	0.014	0.397	0.014	0.334	0.013	0.362	0.024	0.517	0.173	0.532	0.192
EnergyEfficiency_H	0.290	0.011	0.400	0.016	0.406	0.017	0.392	0.017	0.496	0.023	0.315	0.018	0.497	0.191	0.514	0.211
forestFires	0.842	0.528	0.830	0.541	0.831	0.541	0.830	0.541	0.782	0.489	0.815	0.547	0.832	0.529	0.832	0.529
Hitters	0.763	0.131	0.775	0.139	0.782	0.138	0.779	0.139	0.644	0.112	0.718	0.142	0.800	0.159	0.799	0.160
house	0.864	0.024	0.864	0.024	0.864	0.024	0.864	0.024	0.859	0.024	0.867	0.025	0.897	0.044	0.897	0.045
housing	0.543	0.051	0.557	0.052	0.559	0.052	0.562	0.050	0.544	0.055	0.546	0.054	0.639	0.122	0.647	0.128
machineCPU	0.441	0.129	0.441	0.128	0.441	0.128	0.445	0.129	0.384	0.106	0.424	0.154	0.603	0.268	0.634	0.296
mortgage	0.039	0.004	0.040	0.004	0.040	0.004	0.04	0.004	0.038	0.005	0.115	0.007	0.361	0.246	0.372	0.249
mv	0.432	0.004	0.432	0.004	0.432	0.004	0.432	0.004	0.433	0.004	0.440	0.003	0.592	0.163	0.609	0.173

Table A-2. The average and standard deviation of prediction errors of each variable selection techniques for each dataset(part 2)

Data set	Original		Forward		Backward		Stepwise		GA		Ridge		Lasso		Elastic Net	
	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.
nuclear	0.598	0.347	0.679	0.374	0.625	0.378	0.657	0.360	0.518	0.348	0.625	0.650	0.727	0.438	0.736	0.446
OVCTT_bitrate	0.977	0.006	0.987	0.006	0.987	0.006	0.987	0.006	1.237	0.005	0.977	0.006	0.984	0.009	0.984	0.009
OVCTT_framerate	0.990	0.003	0.994	0.003	0.994	0.003	0.994	0.003	1.850	0.003	0.990	0.003	0.993	0.004	0.993	0.004
OVCTT_height	0.849	0.005	0.850	0.005	0.85	0.005	0.850	0.005	0.850	0.004	0.850	0.005	0.877	0.036	0.882	0.040
OVCTT_width	0.846	0.005	0.848	0.006	0.848	0.006	0.848	0.006	0.896	0.006	0.847	0.005	0.875	0.037	0.880	0.040
ParkinsonsTele	0.862	0.016	0.861	0.016	0.861	0.016	0.861	0.016	0.860	0.021	0.861	0.016	0.893	0.039	0.895	0.041
PersonalLoan	0.234	0.010	0.234	0.010	0.234	0.010	0.234	0.010	0.233	0.011	0.233	0.010	0.249	0.021	0.250	0.022
PCASP	0.848	0.004	0.848	0.004	0.848	0.004	0.848	0.004	0.846	0.003	0.854	0.004	0.904	0.045	0.896	0.044
pole	0.731	0.007	0.731	0.007	0.731	0.007	0.731	0.007	55.532	48.985	0.732	0.007	0.813	0.089	0.817	0.091
puma32h	0.885	0.021	0.882	0.020	0.882	0.020	0.882	0.020	1.214	0.190	0.885	0.020	0.908	0.035	0.904	0.036
SML2010	0.341	0.012	0.405	0.013	0.468	0.110	0.405	0.013	0.339	0.011	0.362	0.011	0.505	0.158	0.490	0.174
solar flare_C	0.411	0.058	0.414	0.062	0.414	0.064	0.187	0.034	0.187	0.051	0.393	0.059	0.407	0.061	0.407	0.061
solar flare_M	0.508	0.170	0.509	0.171	0.508	0.170	0.218	0.035	0.232	0.034	0.492	0.171	0.508	0.167	0.507	0.166
solar flare_X	0.129	0.063	0.123	0.070	0.122	0.072	0.123	0.070	0.111	0.080	0.115	0.076	0.126	0.067	0.125	0.067
stock	2.338	0.121	2.335	0.121	2.335	0.121	2.339	0.122	2.723	0.113	2.423	0.097	3.665	1.194	3.766	1.300
student	2.273	0.290	2.357	0.288	2.421	0.283	2.381	0.292	2.286	0.409	2.180	0.280	2.373	0.398	2.380	0.405
tic	0.979	0.036	0.98	0.037	0.981	0.037	0.980	0.037	26.332	57.648	0.976	0.037	0.981	0.037	0.981	0.037
ToyotaCorolla	0.438	0.258	0.548	0.356	0.470	0.261	0.560	0.375	0.310	0.037	0.320	0.023	0.497	0.200	0.503	0.210
UScrime	0.823	0.333	0.849	0.299	0.854	0.289	0.875	0.296	0.687	0.214	0.609	0.201	0.756	0.233	0.750	0.234
Wage	0.299	0.025	0.299	0.025	0.299	0.025	0.299	0.025	0.302	0.030	0.318	0.029	0.484	0.196	0.521	0.207
wankara	0.103	0.004	0.103	0.004	0.103	0.004	0.103	0.004	0.141	0.009	0.139	0.006	0.382	0.240	0.400	0.252
winequality	0.735	0.019	0.737	0.019	0.736	0.019	0.736	0.019	0.731	0.021	0.736	0.019	0.767	0.039	0.770	0.041
wizmir	0.089	0.007	0.089	0.007	0.089	0.007	0.089	0.007	0.091	0.005	0.124	0.009	0.372	0.244	0.398	0.256
wpbc	0.887	0.110	0.878	0.070	0.903	0.073	0.879	0.076	0.843	0.073	0.830	0.079	0.880	0.086	0.878	0.088
yacht	0.583	0.053	0.571	0.052	0.571	0.052	0.571	0.052	0.572	0.051	0.570	0.070	0.668	0.148	0.668	0.149

Table B-1. The average and standard deviation of the proportion of the selected variables by each variable selection techniques for each dataset(part 1)

Data set	Forward		Backward		Stepwise		GA		Ridge		Lasso		Elastic Net	
	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.
abalone	0.875	0.000	0.875	0.000	0.933	0.063	0.825	0.121	1.000	0.000	0.850	0.076	0.883	0.032
akbilgic	0.424	0.026	0.424	0.026	0.424	0.026	0.629	0.138	0.729	0.266	0.476	0.102	0.629	0.071
ANACALT	0.519	0.070	0.519	0.070	0.519	0.070	0.714	0.202	0.748	0.186	0.486	0.221	0.681	0.061
autoMPG8	0.429	0.000	0.429	0.000	0.429	0.000	0.729	0.125	0.952	0.087	0.610	0.091	0.729	0.044
baseball	0.369	0.034	0.373	0.042	0.371	0.040	0.738	0.082	0.788	0.223	0.700	0.125	0.752	0.102
Buzz	0.515	0.050	0.460	0.032	0.469	0.039	0.800	0.058	0.726	0.242	0.053	0.007	0.132	0.011
california	1.000	0.000	1.000	0.000	1.000	0.000	0.963	0.060	1.000	0.000	0.958	0.060	1.000	0.000
Carseats	0.636	0.000	0.636	0.000	0.727	0.000	0.790	0.088	0.808	0.163	0.692	0.079	0.761	0.072
CBM_C	0.902	0.035	0.902	0.035	0.902	0.035	0.929	0.000	0.995	0.018	0.681	0.036	1.000	0.000
CBM_T	0.926	0.013	0.895	0.036	0.895	0.036	0.857	0.000	0.969	0.045	0.686	0.044	0.983	0.041
compactiv	0.646	0.030	0.646	0.030	0.659	0.025	0.886	0.078	0.860	0.112	0.573	0.015	0.610	0.038
concrete	0.746	0.023	0.688	0.064	0.625	0.000	0.888	0.092	0.996	0.023	0.825	0.062	0.883	0.046
dee	0.811	0.058	0.811	0.058	0.833	0.000	0.900	0.086	0.928	0.084	0.833	0.000	0.833	0.000
delta_elv	0.833	0.000	0.833	0.000	0.833	0.000	0.933	0.086	0.917	0.129	0.667	0.000	0.722	0.080
elevators	0.619	0.019	0.619	0.019	0.633	0.043	0.800	0.054	0.904	0.058	0.720	0.045	0.872	0.030
EnergyEfficiency_C	0.567	0.063	0.625	0.000	0.554	0.063	0.875	0.000	0.904	0.097	0.375	0.000	0.525	0.076
EnergyEfficiency_H	0.675	0.062	0.675	0.062	0.550	0.062	0.875	0.000	0.925	0.078	0.404	0.054	0.625	0.033
forestFires	0.083	0.000	0.083	0.000	0.083	0.000	0.433	0.146	0.619	0.356	0.461	0.408	0.461	0.406
Hitters	0.307	0.084	0.311	0.068	0.302	0.070	0.468	0.128	0.748	0.152	0.452	0.221	0.553	0.188
house	0.817	0.023	0.825	0.038	0.925	0.030	0.850	0.067	0.948	0.049	0.917	0.030	0.923	0.027
housing	0.695	0.133	0.690	0.133	0.649	0.098	0.900	0.109	0.931	0.068	0.874	0.145	0.897	0.123
machineCPU	0.794	0.084	0.794	0.084	0.828	0.178	0.717	0.158	0.911	0.150	0.717	0.078	0.767	0.094
mortgage	0.858	0.067	0.789	0.061	0.787	0.069	0.847	0.077	0.940	0.064	0.267	0.000	0.467	0.000
mv	0.500	0.000	0.500	0.000	0.537	0.127	0.830	0.095	0.920	0.113	0.300	0.000	0.400	0.000

Table B-2. The average and standard deviation of the proportion of the selected variables by each variable selection techniques for each dataset(part 2)

Data set	Forward		Backward		Stepwise		GA		Ridge		Lasso		Elastic Net	
	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.	Avg.	Sdev.
nuclear	0.573	0.131	0.600	0.053	0.567	0.071	0.560	0.178	0.667	0.278	0.717	0.238	0.740	0.246
OVCTT_bitrate	0.176	0.000	0.176	0.000	0.176	0.000	0.467	0.000	0.556	0.272	0.261	0.059	0.278	0.058
OVCTT_framerate	0.176	0.000	0.176	0.000	0.176	0.000	0.667	0.000	0.357	0.264	0.228	0.051	0.246	0.063
OVCTT_height	0.237	0.011	0.237	0.011	0.245	0.031	0.733	0.000	0.880	0.156	0.267	0.023	0.272	0.017
OVCTT_width	0.239	0.015	0.241	0.024	0.251	0.038	0.600	0.000	0.889	0.153	0.267	0.023	0.274	0.014
ParkinsonsTele	0.553	0.018	0.550	0.000	0.550	0.032	0.785	0.078	0.873	0.117	0.665	0.044	0.712	0.022
PersonalLoan	0.561	0.037	0.561	0.037	0.561	0.037	0.758	0.100	0.950	0.068	0.767	0.077	0.778	0.091
PCASP	1.000	0.000	1.000	0.000	1.000	0.000	0.944	0.108	1.000	0.000	0.944	0.057	1.000	0.000
pole	0.454	0.033	0.554	0.042	0.456	0.032	0.769	0.073	0.999	0.007	0.591	0.024	0.710	0.028
puma32h	0.038	0.013	0.038	0.013	0.038	0.013	0.359	0.123	0.403	0.186	0.257	0.206	0.372	0.193
SML2010	0.166	0.009	0.175	0.029	0.168	0.010	0.748	0.076	0.949	0.036	0.427	0.201	0.544	0.034
solar flare_C	0.122	0.057	0.113	0.057	0.120	0.053	0.400	0.000	0.637	0.276	0.441	0.342	0.462	0.361
solar flare_M	0.098	0.036	0.128	0.065	0.108	0.072	0.300	0.000	0.768	0.176	0.443	0.326	0.456	0.319
solar flare_X	0.063	0.045	0.097	0.052	0.068	0.046	0.200	0.000	0.376	0.392	0.265	0.344	0.270	0.336
stock	0.889	0.000	0.889	0.000	0.989	0.045	0.967	0.054	1.000	0.000	0.778	0.000	0.841	0.056
student	0.195	0.038	0.156	0.040	0.166	0.020	0.540	0.189	0.860	0.096	0.625	0.190	0.646	0.190
tic	0.126	0.018	0.120	0.016	0.122	0.011	0.313	0.037	0.586	0.216	0.503	0.187	0.511	0.180
ToyotaCorolla	0.435	0.068	0.457	0.061	0.437	0.062	0.736	0.120	0.836	0.086	0.450	0.098	0.514	0.070
UScrime	0.247	0.111	0.276	0.099	0.278	0.096	0.253	0.117	0.927	0.088	0.518	0.267	0.616	0.239
Wage	0.176	0.000	0.176	0.000	0.176	0.000	0.711	0.119	0.996	0.014	0.102	0.021	0.111	0.000
wankara	0.889	0.000	0.889	0.000	0.893	0.020	0.967	0.075	0.856	0.134	0.341	0.028	0.556	0.000
winequality	0.621	0.107	0.718	0.134	0.697	0.127	0.855	0.098	0.948	0.052	0.764	0.051	0.785	0.056
wizmir	0.622	0.055	0.622	0.055	0.667	0.000	0.800	0.126	1.000	0.000	0.222	0.000	0.437	0.028
wpbc	0.128	0.021	0.134	0.032	0.135	0.019	0.242	0.055	0.711	0.164	0.566	0.261	0.642	0.241
yacht	0.167	0.000	0.167	0.000	0.167	0.000	0.350	0.200	0.756	0.262	0.167	0.000	0.239	0.095