理學碩士學位請求論文

랜덤포레스트를 이용한 변수 선택

Variable Selection Using Random Forest

2013年 2月

仁荷大學校 大學院

統計學科 統計學專攻

권 안 나

理學碩士學位請求論文

랜덤포레스트를 이용한 변수 선택

Variable Selection Using Random Forest

2013年 2月

指導敎授 朴 憲 鎭

이 論文을 碩士學位 論文으로 提出함

仁荷大學校 大學院

統計學科 統計學專攻

권 안 나

이 論文을 권안나의 碩士學位論文으로 認定함.

2013年 2月

主審 _____

副審 _____

委員 _____

요 약

본 연구에서는 데이터 마이닝이나 빅데이터에서 이슈가 되고 있는 변수 선택 문제를 해결하기 위한 다양한 방법을 비교 연구하였다. 지금까지 주로 연구에서 사용하는 변수 선택 방법에는 상관분석, R-Square, Adjust R-Square, stepwise 등 회귀분석 기법을 이용한 방법과 의사결정나무와 의사결정나무에 앙상블 기법을 적용한 Bagging, Bumping 등의 방법이 알려져 있다. 이들 방법과 더불어 의사결정나무 개념을 적용하여 발전된 랜덤 포레스트의 방법을 이용하여 변수를 선택하는 방법을 제시하고자 한다.

기존의 방법들과 랜덤 포레스트의 결과를 비교하기 위하여 프리드만(1984)의 모형, 장영재(2008)의 모형, 기타 모형을 이용한 시뮬레이션 자료를 생성하고, 결과를 비교한다. 그 결과, 랜덤 포레스트는 다른 방법들에 비해 주변수를 선택하는 기능은 약하지만 noise 변수를 걸러내는 데에는 효과를 나타낸다.

다음으로는 Cellulose 용해에 영향을 주는 최적 용매를 랜덤포레스트를 이용하여 선택하고, 선택된 변수를 이용하여 모형에 적합하여 예측한 결과와 전체 변수를 이용하여 예측한 결과를 비교한다. 비교 결과, 랜덤 포레스트를 이용하여 선택한 변수를 모형에 적합했을 때가 전체 변수를 이용한 모형보다 성능이 더 좋게 나타난다.

또한 bagging, 랜덤 포레스트, 회귀분석으로 선택된 변수들을 이용하여 MARS 분석을 실시하고, 전체 변수를 이용하여 분석한 결과와 비교한 결과, 선택된 변수들을 가지고 분석한 모형의 성능이 더 좋게 나타난다.

Abstract

In this paper, we made a comparative study for solving a variable selection problem became an issue in data mining or big data. We have used many variable selection methods. For example, there are regression analysis techniques such as correlation, R-square, Adjust R-Square, stepwise and there are bagging method and bumping method that applying ensemble technique for a decision tree. In addition, we proposed a variable selection method using random forest developed from decision tree.

For comparing random forest with previous methods, we generated simulation data based on the Friedman(1984), the Jang Yeong-jae (2008), and others. And we compared the performance of random forest with those of some other variable selection methods. Consequently, random forest is effective for eliminating noise variables but it is weak for selecting main variables compared with other variable selection methods.

Secondly, we used a data about the dissolution of cellulose. In this data, we extracted some optimum variables by random forest. After that, we compared the prediction performance of the model using selected variables with that of another using all. Consequently, when we fitted a regression model and a tree model with some main variables selected from random forest, the results appeared to be better than using all variables.

Finally, we progressed MARS analysis using selected variables by

bagging, random forest, regression analysis and compared the results using selected variables with that of using all variables. Consequently, the results appeared to be better when we used selected variables.

목 차

1. 서 론

오늘날, 정보력의 발달로 인해 그에 관련된 데이터의 양이 기하급수적으로 증가하면서 어떠한 문제에 대한 결과를 예측하는 데에 영향을 미치는 요인이 많아지고 있다. 이러한 big data가 가지는 수많은 설명변수들 중에서 가장 영향력 있고 예측력 있는 변수를 선택하여 데이터의 양을 줄이는 것은 금융, 제조, 서비스 등 여러 분야에서 중요한 과정으로 생각되고 있다.

결과를 예측하는 데에 필요한 모형 구축을 위해 현재는 대부분 회귀분석이나 의사결정나무 등의 분석을 통해 중요한 변수를 선택하고, 그 결과를 이용하여 모형 구축에 활용하고 있다. 하지만 회귀분석의 경우 설명변수와 반응변수가 선형적인 관계에 있는 모형에서 변수 선택 비율이 높은 편이지만 비선형적인 관계에 있는 모형에서는 변수 선택 비율이 낮아진다는 단점이 있다. 의사결정나무의 경우에는 다른 방법들에 비하여 설명하기가 쉽다는 장점과 변수의 개수에 영향을 덜 받는다는 장점이 있기는 하지만 결과가 초기의 분할에 큰 영향을 받게 되고, 변수간의 교호작용이 지나치게 강조되며 분석 자료의 최대값이나 최소값의 범위를 벗어나는 값은 예측이 어렵다는 단점이 있다.

기존 방법들보다 더 나은 모형 적합을 위하여 프리드만(1984), 장영재(2008), 기타 모형 등 다양한 상황에 맞는 모형을 이용하여 랜덤 포레스트 분석을 통해 뽑히는 횟수가 적은 노이즈 변수를 걸러내고 모형에 필요한 주변수를 선택한 후, 다른 방법들과 결과를 비교해보고자 한다.

2. 연구배경

데이터마이닝에서는 회귀분석과 의사결정나무 모형을 이용하여 분석을 하여 변수를 선택하고 모델링을 하고 있다. 지금까지의 변수 선택법에서의 변수 선택의 어려움을 살펴보고, 랜덤 포레스트의 반복성을 이용하여 선택되는 변수들의 확률을 이용하여 변수를 선택해보려고 한다.

2.1 상관분석

상관분석은 반응변수와 설명변수들 간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법으로서 두 변수간의 상관관계의 정도를 상관계수(Correlation Coefficient)값을 구하여 나타내는 방법이다. 상관계수 값은 아래와 같이 구한다.

$$\rho = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

마이닝 분석에서 상관분석을 이용하여 변수선택을 할 경우에 주로 상관계수 값의 제곱 값이 기준 값보다 작을 때에 변수에서 제거하는 방법을 이용하고 있다. 이 때 기준 값은 보통 0.005를 많이 사용한다.

하지만 상관분석에서 사용하는 상관계수는 두 변수간의 선형적인 관계에 중점을 두고 분석을 하기 때문에 변수들 사이에 비선형적인 관계가 있는 경우에는 상관계수의 값이 작게 나타나게 되고, 실제로는 중요한 변수임에도 분석에서는 선택되지 않는 결과를 나타낼 수 있다.

2.2 회귀분석

회귀분석은 둘 또는 그 이상의 변수들 간에 존재하는 관련성을 분석하기 위하여 관측된 자료에서 이들 간의 함수적 관계식을 각기 다른 가중치를 곱한 선형함수로 나타내어 반응변수를 통계적 방법으로 추정하는 방법이다. 회귀분석에서의 대표적인 변수 선택법에는 전진 선택법(Forward Selection), 후진 제거법(Backward Elimination), 단계 선택법(Stepwise Selection) 등이 있어서 이 중의 한 가지 방법으로 변수를 선택하여 모형을 구축하게 된다.

전진 선택법(Forward Selection)은 변수의 개수를 1개부터 시작하여 수많은 설명변수들 중에 반응변수에 가장 크게 영향을 줄 것으로 판단되는 변수부터 하나씩 선택하여 더 이상 중요한 변수가 없다고 판단될 때 변수의 선택을 중단하는 방법이다.

후진 제거법(Backward Elimination)은 변수 전체를 시작으로 하여 수많은 설명변수들 중에 반응변수에 가장 작게 영향을 줄 것으로 판단되는 변수부터 하나씩 제거하면서 더 이상 제거할 변수가 없다고 판단될 때 변수의 제거를 중단하는 방법이다.

단계 선택법(Stepwise Selection)은 전진 선택법과 후진 제거법의 단점을 보완한 방법으로 중요한 변수를 하나씩 선택하여 나가면서 이미 선택된 변수에 새로운 변수가 추가되면서 중요성을 상실하여 제거될 수 있는지를 매 단계별로 검토하는 방법이다. 위의 절차를 계속 밟아가며 새로이 선택된 변수가 유의하지 않을 때까지 계속 진행된다.

회귀분석을 이용한 변수 선택법은 선형인 관계를 갖는 모형에서는 효과적일 수 있으나 비선형의 관계를 갖는 모형에서는 적용할 수 없다는 문제점을 가지고 있다.

2.3 의사결정나무

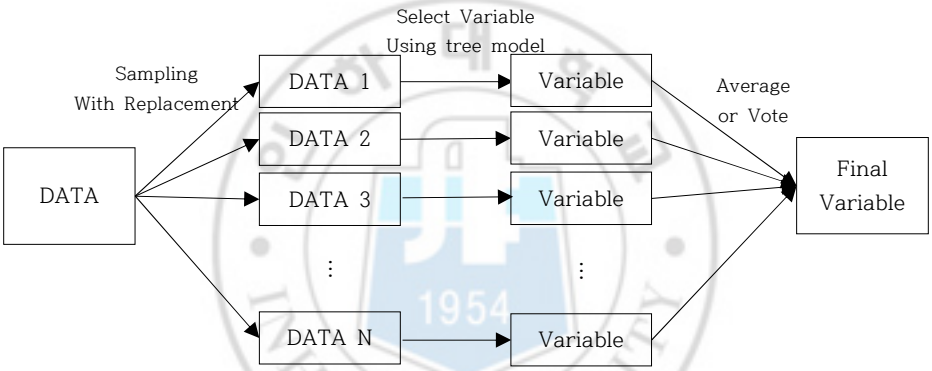
의사결정나무 모형은 주어진 변수의 규칙 혹은 조건문을 토대로 나무 구조로 도표화하여 분류와 예측을 목적으로 수행하는 방법이다. 데이터를 비슷한 특성을 갖는 몇 개의 그룹으로 분할하여 각 그룹별 특성을 발견하여 한 데이터가 어떤 그룹에 속하게 되는지 세분화하는 경우, 데이터를 여러 변수들에 근거하여 목표변수의 범주를 몇 개의 등급으로 분류하는 경우, 데이터로부터 규칙을 찾아내고 이를 이용하여 미래의 사건을 예측하는 경우, 매우 많은 수의 변수 중에서 반응변수에 큰 영향을 미치는 변수들을 선택하는 경우, 여러 개의 설명변수들이 결합하여 반응변수에 작용하는 규칙이나 교호작용효과를 파악하는 경우 등에 많이 이용하고 있다.

의사결정나무는 분류 또는 예측의 과정이 나무구조로 도표화하여 표현되기 때문에 다른 방법들(회귀분석 등)에 비하여 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다. 또한 매우 많은 수의 변수 중에서 반응변수에 큰 영향을 미치는 변수들을 선택하는 경우에 사용되고 있는 만큼, 변수의 개수에 영향을 덜 받는 편이고 다양한 형태의 변수에 적용이 가능하여 변수의 중요도를 파악하는데 효과적이고 해석이 용이하다.

하지만 데이터의 개수에 민감하여 재귀적인 알고리즘에 의해 결과가 초기의 분할에 큰 영향을 받게 되고, 상대적으로 많은 분리가 가능한 변수로 분리 기준이 정해지는 현상이 있다. 또한 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근방에서는 예측 오류가 클 가능성이 크며, 특정 모형들에 대해서 분할이 어려워 변수 선택이 유의하게 되지 않을 가능성이 있다는 문제점이 있다.

2.4 Bagging

데이터가 조금이라도 바뀐 상태에서 분류자의 변동성이 큰 경우에는 예측자의 변동성을 감소시키고자 부스트랩(bootstrap) 방법을 통해 분류자를 얻을 수 있다. 이러한 방법을 bagging(Bootstrap Aggregating) 알고리즘이라 하며, bagging은 부스트랩 방법을 이용한 앙상블 기법으로 Leo Breiman(1996)에 의해 소개되었다.



[그림2.1] Bagging 알고리즘 과정

[그림2.1]의 bagging 알고리즘 과정은 모집단으로부터 추출된 training dataset에서 복원 추출에 의해 부스트랩 데이터를 생성한다. 이러한 방법을 N번 반복하여 N개의 부스트랩 데이터를 생성하고, 각각의 부스트랩 데이터에 적합한 분류 알고리즘(의사결정나무 등)을 적용하여 각각의 단일 분류자를 형성하여 N개의 단일 분류자 집합을 얻는다. 이러한 단일 분류자를 결합하는 방법에는 반응 변수가 연속형일 때에는 평균, 범주형일 때에는 투표를 사용한다. 이렇게 결합되어 형성된 분류자를 bagging 분류자라고 한다.

Breiman에 의하면 training dataset이 불안정하다면 bagging 분류자의 결합을 통해서 분류 성능이 향상되어진다고 한다. 그러나 training dataset이 안정적이라면 bagging 과정을 통해서 얻어진 bagging 분류자는 training dataset에서 얻어진 단일 분류자와 비슷하다고 한다.

Bagging은 모형의 분류 정확성을 높이고 안정성을 향상시키는 효과가 있다. 더불어 분산을 줄이고, overfitting을 피하는데 도움을 준다. Bagging은 대개 의사결정나무 모형에서 많이 사용됨에도 불구하고 다른 모형에서도 적용이 가능하다는 장점이 있다.

2.5 Bumping

Bumping방법론은 모델의 결합 또는 평균치로 나타내는 방법이 아닌 가장 좋은 하나의 모델을 찾아내는 방법이다. Bumping은 랜덤하게 생성되는 부스트랩 표본(bootstrap sampling)을 이용한다. 또한 bumping은 수많은 local minima에 빠져있는 문제를 해결하는 곳에 많이 쓰이고 모델을 찾아내기 힘든 모형에서의 해결할 수 있는 하나의 방법으로 쓰인다.

Bagging방법론에서와 같이 부스트랩 표본을 이용하여 각각의 표본에 모델을 적합시킨다.

그러나 bagging방법론처럼 performance(prediction error)를 평균 내는 것이 아니고 부스트랩 표본 중에 최적의 performance로 예측되는 모형을 선택하게 된다. 즉, 부스트랩 표본 Z_1, ..., Z_N이 있고 각각에 모델을 적합하게 되며, 예측 값이 $\widehat{f^N}(x)$, N=1,2,3,...,n이 나타난다. 이 중에서 가장 작은 prediction error를 나타내는 모델을 선택하게 된다.

예를 들어, 부스트랩 표본 $\hat{b}$에서 모델을 선택하게 되고, 그 때 이런 수식으로 표현이 된다.

$$\hat{b} = \underbrace{\mathrm{argmin}}_{n} \sum_{i=1}^{N} [y_i - \hat{f}^n(x_i)]^2$$

모델의 예측 값은 $\hat{f}^N(x)$이고 본래의 training dataset을 이용하여 각각의 performance인 prediction error를 구하게 된다. Bumping은 모델공간에서 가장 좋은 적합력을 보여주는 모델을 찾는 방법으로 패턴을 찾기 힘든 데이터에서 사용하기 좋은 방법이다. 예를 들어, 미세한 데이터 부분이 패턴을 찾는 데에 있어서 방해를 하고 있다. 그럴 때에 부스트랩 표본을 이용한 bumping 방법론을 이용한다면 더 좋은 결과를 가져올 수 있다.

2.6 Random Forest

2.4절에서 본 bagging 방법의 핵심은 부스트랩 데이터의 분류 알고리즘 결과를 평균내거나 투표를 통해 예측을 한다는 것이다. 이러한 bagging 방법에 임의의 additional layer를 추가한 방법이 랜덤 포레스트이고, Leo Breiman(2001)에 의해 소개되었다.

기존에 연구되어온 의사결정나무에서는 모든 변수를 사용하여 가장 최적의 결과를 내는 분할로 각각의 노드(node)를 나타낸 것과 달리 랜덤 포레스트에서는 각각의 노드를 나타낼 때, 설명변수를 무작위로 선택하고, 선택된 설명변수의 집합 중에서 가장 최적의 결과를 내는 방법을 이용한다.

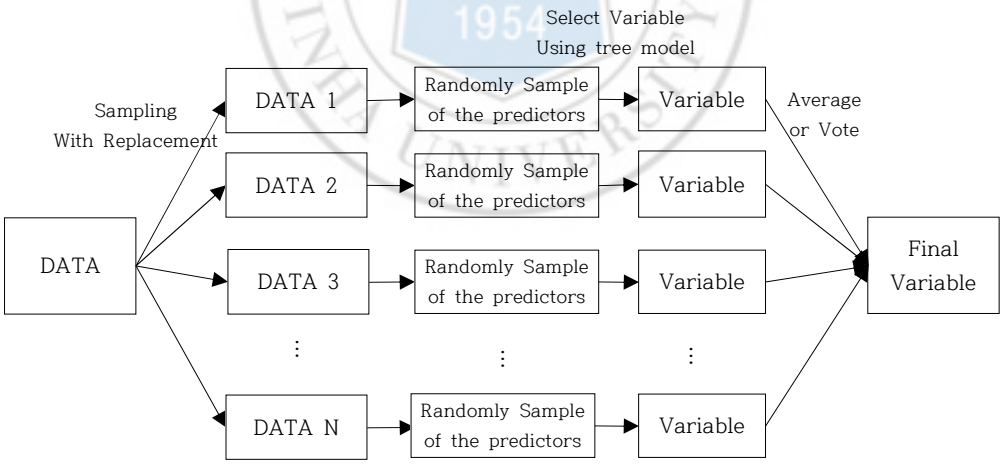
[그림2.2]의 랜덤 포레스트 알고리즘 과정은 모집단으로부터 추출된 training dataset에서 복원 추출에 의해 부스트랩 데이터를 생성한다. 이러한 방법을 N번 반복하여 N개의 부스트랩 데이터를 생성하고, 의사결정나무 알고리즘을 적용할 때 각각의 노드에서 랜덤하게 m개의 설명변수를

선택한다. (전체 설명변수를 M개라고 할 경우, Classification Tree에서는 보통 m= $\sqrt{M}$ 개의 변수를 선택하고, Regression Tree에서는 보통 m=M/3 개의 변수를 랜덤으로 선택하여 적용한다.). 이렇게 선택된 변수들을 이용하여 가장 최적의 split 조합을 찾아서 각각 N개의 단일 분류자 집합을 얻고, bagging 방법론(bagging 방법론은 랜덤 포레스트 방법론에서 선택하는 변수의 개수가 m=M인 특별한 경우가 된다.)과 마찬가지로 반응변수가 연속형인 경우에는 평균, 범주형인 경우에는 투표를 사용하여 단일 분류자를 결합한다.

Regression : $\widehat{f_{rf}^{N}}(x) = \frac{1}{N}\sum_{b=1}^{N} T_b(x)$ (단, $T_b(x)$: random forest tree)

Classification : $\widehat{C_{rf}^{N}}(x) = majority\ vote\ \widehat{C}_b(x)_1^N$

(단, $\widehat{C}_b(x)$: the class prediction of the b-th random forest tree)



[그림2.2] 랜덤 포레스트 알고리즘 과정

랜덤 포레스트 방법론은 트리(tree) 사이에 상관관계를 줄임으로써 bagging 방법에 비해 분산을 줄여준다는 장점을 가지고 있다. 또한 기존

의 다른 알고리즘에 비해 정확한 결과를 나타내며 변수를 제거하지 않고, 수천 개의 독립 변수를 모두 활용할 수 있기 때문에 대형자료에서 중요한 변수를 찾는데 용이하다. 특히 입력변수의 개수가 많을 때에는 bagging이나 boosting과 비슷하거나 더 좋은 예측력을 보이는 경우가 많은 것으로 알려져 있다. 하지만 아직은 랜덤 포레스트에 대한 이론적 설명이나 최종 결과에 대한 해석이 어렵다는 단점을 가지고 있다.

랜덤 포레스트에 대한 본 연구는 bagging 방법을 이용한 변수 선택법에서 시작되었다. Bagging을 이용한 변수 선택법과 마찬가지로 부스트랩 방법을 통하여 분류자를 얻을 수 있고, 각각의 노드를 분할할 때, 전체의 설명변수를 대상으로 분할할 수 있는 변수를 찾는 것이 아니라 랜덤으로 일정 개수의 변수를 선택한 후에 그 중에서 가장 최적으로 분할할 수 있는 변수를 활용하는 것이다. 또한 bagging tree의 방법을 확장하여 하나의 데이터 셋을 가지고 n개의 트리를 생성한 후, 각각의 트리에서 뽑힌 변수들의 횟수를 통해 변수의 중요도를 파악할 수 있고, 이를 N개의 데이터 셋으로 확장하여 선택된 변수들을 조합한다면 여러 번의 시행을 통해 선택된 변수들에 대한 신뢰성을 높일 수 있다.

본 연구에서는 주어진 원데이터를 가지고 복원추출을 이용하여 N개의 부스트랩 데이터 셋을 형성하고, 랜덤 포레스트의 과정을 거쳐 변수를 선택하게 된다. 위의 과정을 N번 반복하여서 선택된 변수들의 횟수를 파악하고, 적게 뽑히는 노이즈 변수는 제거하고, 많이 뽑히는 변수를 찾아 주변수로 이용하고자 한다.

3. 시뮬레이션을 통한 변수 선택 결과 비교

기존에 알려진 다른 방법들과 비교하며 랜덤 포레스트를 이용한 변수 선택 결과의 정확성을 측정하기 위해 시뮬레이션을 시행하였다. 본 시뮬레이션에서는 프리드만(1984), 장영재(2008), Saddle 모형으로 랜덤 포레스트를 수행하여 다양한 변수들 중에서 우리가 관심을 가지는 주변수들이 얼마나 정확하게 선택되는지와 noise 변수에 대한 결과를 살펴보았다.

3.1 프리드만의 모형

3.1절에서는 프리드만(1984)에서 사용된 5가지 함수를 사용하였고, 모형은 $y = f(x_1, x_2) + 0.25\epsilon$이다. 이 때, x_1과 x_2는 0과 1사이의 균일분포를 따르고, ϵ은 평균이 0이고 분산이 1인 표준정규분포를 따른다고 가정하였다. 이러한 분포를 가진 데이터를 랜덤하게 1000개 생성하여 분석에 이용하였다. 또한 표준정규분포를 따르는 noise 변수를 10개, 0과 1 사이의 균일분포를 따르는 noise 변수를 10개 추가로 만들고 그에 따른 데이터를 랜덤하게 1000개 생성하여 다양한 분석 방법을 통해 주변수인 x_1, x_2가 얼마나 많이 선택되고, noise 변수가 얼마나 적게 선택되는지를 살펴보았다. 시뮬레이션에 사용된 5가지의 함수는 아래와 같고, 모형은 [그림3.1]과 같다.

① Simple Function

$f(x_1, x_2) = 10.391((x_1 - 0.4)(x_2 - 0.6) + 0.36)$

② Radial Function

$f(x_1, x_2) = 24.234(r^2(0.75 - r^2))$, with $r^2 = (x_1 - 0.5)^2 + (x_2 - 0.5)^2$

③ Harmonic Function

$$f(x_1,x_2) = 42.659(0.1 + x_1^*(0.05 + x_1^{*4} - 10x_1^{*2}x_2^{*2} + 5x_2^{*4}))$$

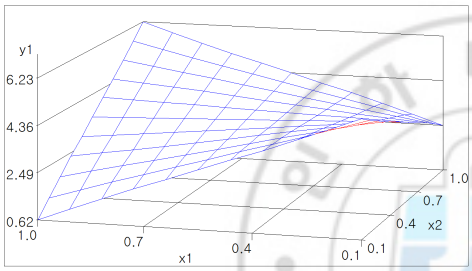
with $x_1^* = x_1 - 0.5,\ x_2^* = x_2 - 0.5$

④ Additive Function

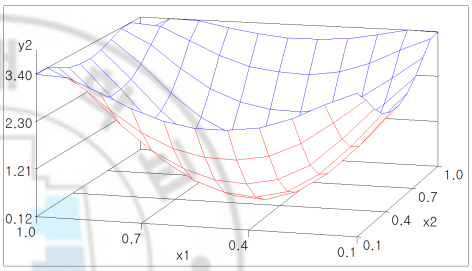
$$f(x_1,x_2) = 1.3356(1.5(1-x_1) + e^{2x_1-1}\sin(3\pi(x_1-0.6)^2) + e^{3(x_2-0.5)}\sin(4\pi(x_2-0.9)^2))$$

⑤ Complicated Function

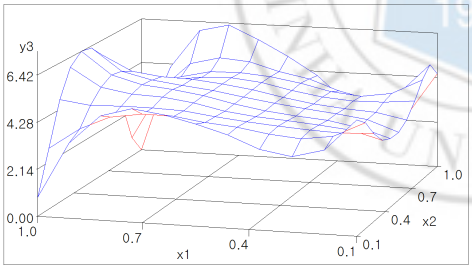
$$f(x_1,x_2) = 1.9(1.35 + e^{x_1}\sin(13(x_1-0.6)^2)e^{-x_2}\sin(7x_2))$$



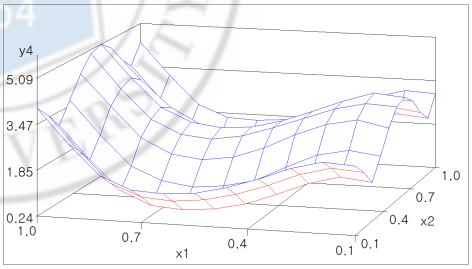
Simple Function



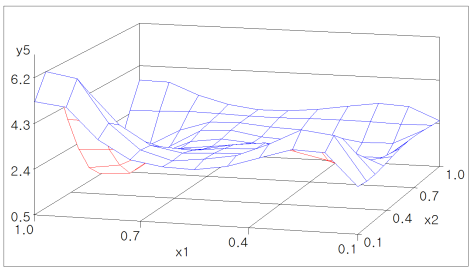
Radial Function



Harmonic Function



Additive Function



Complicated Function

[그림3.1] 프리드만(1984)의 모형

원 데이터에서 복원추출을 이용하여 100개의 부스트랩 데이터 셋을 형성하여 랜덤 포레스트를 이용해 어떤 변수가 주변수로 선택되는지 확인하였다. 각각의 모형에 대해 100번의 실험 결과 주변수로 선택된 변수의 선택 비율은 아래 [표3.1]과 같다. [표3.1]에서 보면 변수 선택법으로 랜덤 포레스트를 사용했을 때, 각각의 모형에서 주변수로 사용되었던 x_1과 x_2가 거의 90% 이상의 비율로 선택되고, noise 변수로 불리는 나머지 변수들은 낮은 평균값을 보였다.

특히 기존에 알려진 변수선택법을 이용한 결과와 비교했을 때, Radial 모형에서 회귀분석, R-square 등의 방법에 비해 트리 계열(트리, Bagging Tree, 랜덤 포레스트 등)을 사용했을 때, 주변수가 훨씬 더 많이 선택되는 것을 볼 수 있다.

[표3.1]을 noise 변수 관점에서 보면 랜덤 포레스트는 bagging tree 등 기존의 다른 트리 계열에 비해 noise 변수가 선택되는 비율이 낮게 나타나거나 비슷한 것을 알 수 있다. 따라서 랜덤 포레스트 방법을 사용했을 때의 결과는 주변수를 선택하는 면에서는 약하지만 noise 변수를 걸러내는 데에는 다른 방법들에 비해 효과적이라고 할 수 있다.

변수 / 모형	변수선택법	모형 변수 (%)		noise 변수 (%)
		x_1	x_2	
Simple	Corr	100	100	2.35
	R-square	100	100	42.25
	Adj-Rsq	100	100	20.6
	CP	100	100	7.4
	Forward	100	100	4.95
	Backward	100	100	5.15
	Stepwise	100	100	4.95
	Tree	100	100	3.55
	Bagging Tree	100	100	22.79
	Bumping Tree	100	100	24.95
	Random Forest	90.96	90.34	19.56

Radial	Corr	2	1	2.55
	R-square	38	43	46.35
	Adj-Rsq	19	15	22.65
	CP	7	7	7.8
	Forward	4	3	5.8
	Backward	4	3	6
	Stepwise	4	3	5.8
	Tree	100	100	3.15
	Bagging Tree	100	100	22.88
	Bumping Tree	100	100	21.62
	Random Forest	88.1	88.22	18.15
Harmonic	Corr	99	7	2.65
	R-square	100	59	45.9
	Adj-Rsq	100	31	22.35
	CP	99	12	8.15
	Forward	100	11	4.95
	Backward	100	11	4.95
	Stepwise	100	11	4.95
	Tree	100	99	2.85
	Bagging Tree	99.97	98.81	15.05
	Bumping Tree	100	100	15.62
	Random Forest	91.89	68.94	21.37
Additive	Corr	38	100	2.75
	R-square	92	100	47
	Adj-Rsq	74	100	23.1
	CP	54	100	8.2
	Forward	60	100	5.2
	Backward	61	100	5.35
	Stepwise	60	100	5.2
	Tree	100	100	3.35
	Bagging Tree	100	100	19.1
	Bumping Tree	100	100	17.33
	Random Forest	91.64	90.21	17.95
Complicated	Corr	76	100	2.9
	R-square	96	100	43.75
	Adj-Rsq	92	100	21.95
	CP	81	100	8.75
	Forward	88	100	5.15
	Backward	88	100	5.25
	Stepwise	88	100	5.15
	Tree	100	100	2.95
	Bagging Tree	100	100	17.72
	Bumping Tree	100	100	16.47
	Random Forest	87.28	92.83	19.77

[표3.1] 프리드만(1984) 모형의 변수 선택 결과

3.2 장영재의 모형

3.2에서는 장영재(2008)에서 사용된 5가지 함수를 사용하였고, 모형은 $y = f(x_1, x_2, x_3, x_4, x_5) + \epsilon$이다. 이 때, x_1, x_2, x_3, x_4, x_5는 0과 1사이의 균일분포를 따르고, ϵ은 평균이 0이고 분산이 1인 표준정규분포를 따른다고 가정하였다. 이러한 분포를 가진 데이터를 랜덤하게 1000개 생성하여 분석에 이용하였다. 또한 표준정규분포를 따르는 noise 변수를 10개, 0과 1 사이의 균일분포를 따르는 noise 변수를 10개 추가로 만들고 그에 따른 데이터를 랜덤하게 1000개 생성하여 다양한 분석 방법을 통해 주변수인 x_1, x_2, x_3, x_4, x_5가 얼마나 많이 선택되고, noise 변수가 얼마나 적게 선택되는지를 살펴보았다. 시뮬레이션에 사용된 5가지의 함수는 아래와 같다.

① Model 0

$$f(x_1, x_2, x_3, x_4, x_5) = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5$$

② Model 1

$$f(x_1, x_2, x_3, x_4, x_5) = 0.1e^{4x_1} + \frac{4}{1 + e^{-20(x_2 - 0.5)}} + 3x_3 + 2x_4 + x_5$$

③ Model 2

$$f(x_1, x_2, x_3, x_4, x_5) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

④ Model 2a

$$f(x_1, x_2, x_3, x_4, x_5) = 10\sin(4\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

⑤ Model 3

$$f(x_1, x_2, x_3, x_4, x_5) = 10(x_1 + x_2 + x_3 + x_4 + x_5 - 2.5)^2$$

변수 / 모형	변수선택법	모형 변수 (%)					noise 변수 (%)
		x_1	x_2	x_3	x_4	x_5	
Model 0	Corr	93	100	100	100	100	2.2
	R-square	100	100	100	100	100	9.95
	Adj-Rsq	100	100	100	100	100	16.95
	CP	100	100	100	100	100	7.3
	Forward	100	100	100	100	100	5.05
	Backward	100	100	100	100	100	5.25
	Stepwise	100	100	100	100	100	5.05
	Tree	35	100	100	100	100	4.1
	Bagging Tree	62.03	100	100	100	100	26.97
	Bumping Tree	65.01	99.99	100	100	100	26.28
	Random Forest	26.98	58.95	77.97	88.69	92.51	10.19
Model 1	Corr	100	100	100	100	90	2.3
	R-square	100	100	100	100	100	15.4
	Adj-Rsq	100	100	100	100	100	17.7
	CP	100	100	100	100	100	7.75
	Forward	100	100	100	100	100	5.55
	Backward	100	100	100	100	100	5.55
	Stepwise	100	100	100	100	100	5.55
	Tree	100	100	100	99	32	3.7
	Bagging Tree	100	100	100	100	64.68	24.57
	Bumping Tree	100	100	100	100	68.6	24.34
	Random Forest	89.01	92.73	76.71	58.74	28.36	10.99
Model 2	Corr	100	100	1	100	100	2.45
	R-square	100	100	26	100	100	18.15
	Adj-Rsq	100	100	25	100	100	18.25
	CP	100	100	13	100	100	6.55
	Forward	100	100	10	100	100	4.55
	Backward	100	100	10	100	100	4.75
	Stepwise	100	100	10	100	100	4.55
	Tree	100	100	86	100	100	3.75
	Bagging Tree	100	100	92.09	100	100	23.63
	Bumping Tree	100	100	94.5	100	100	22.99
	Random Forest	82.92	82.44	55.93	92.06	71.49	10.86
Model 2a	Corr	100	100	2	100	100	2.55
	R-square	100	100	38	100	100	40.65
	Adj-Rsq	100	100	16	100	100	19.8
	CP	100	100	6	100	100	7.5
	Forward	100	100	2	100	100	5.05
	Backward	100	100	2	100	100	5.05
	Stepwise	100	100	2	100	100	5.05
	Tree	100	100	36	100	86	3.25

	Bagging Tree	100	100	64.88	100	94.18	25.3
	Bumping Tree	100	100	65	100	94	24.28
	Random Forest	80.05	79.55	45.89	89.46	63.86	14.32
Model 3	Corr	8	8	9	6	7	3.2
	R-square	60	65	54	47	56	49.95
	Adj-Rsq	29	36	30	27	33	22.55
	CP	15	14	19	12	10	8.95
	Forward	14	10	14	9	9	5.3
	Backward	14	10	14	9	9	5.5
	Stepwise	14	10	14	9	9	5.3
	Tree	62	60	64	62	63	2.3
	Bagging Tree	75.27	74.72	76.16	76.23	75.51	13.44
	Bumping Tree	97.2	97.1	97.6	98.6	98.2	15.96
	Random Forest	58.55	59.39	59.25	59.82	59.5	19.48

[표3.2] 장영재(2008) 모형의 변수 선택 결과

원 데이터에서 복원추출을 이용하여 100개의 부스트랩 데이터 셋을 형성하여 랜덤 포레스트를 이용해 어떤 변수가 주변수로 선택되는지 확인하였다. 각각의 모형에 대해 100번의 실험 결과는 위의 [표3.2]와 같다. [표3.2]에서 보면 변수 선택법으로 랜덤 포레스트를 사용했을 때, 각각의 모형에서 주변수로 사용되었던 x_1, x_2, x_3, x_4, x_5가 높은 비율로 선택되고, noise 변수로 불리는 나머지 변수들은 낮은 평균값을 보였다.

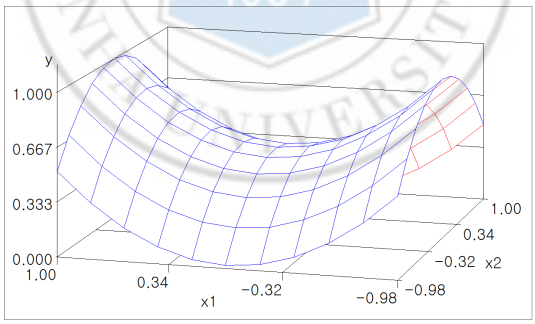
특히 기존에 알려진 변수선택법을 이용한 결과와 비교했을 때, Model 2와 Model 2a에서 회귀분석, R-square 등의 방법에 비해 트리 계열(트리, Bagging Tree, 랜덤 포레스트 등)을 사용했을 때, 주변수인 x_3 변수가 훨씬 더 높은 비율로 선택되는 것을 볼 수 있다. 또한 Model 3에서는 다른 방법에 비해 트리 계열을 사용했을 때, 주변수인 x_1, x_2, x_3, x_4, x_5가 높은 비율로 선택되는 것을 볼 수 있다.

또한 [표3.1]에서와 마찬가지로 [표3.2]를 noise 변수 관점에서 보면 랜덤 포레스트는 기존의 다른 트리 계열에 비해 noise 변수가 선택되는 비율이 훨씬 낮게 나타나거나 비슷한 것을 알 수 있다.

3.3 기타 모형

3.3에에서는 기타 모형으로 말안장의 형태를 보이는 함수를 사용하였고, 모형은 $y = f(x_1, x_2) + 0.25\epsilon$이다. 이 때, x_1과 x_2는 0과 1사이의 균일분포를 따르고, ϵ은 평균이 0이고 분산이 1인 표준정규분포를 따른다고 가정하였다. 이러한 분포를 가진 데이터를 랜덤하게 1000개 생성하여 분석에 이용하였다. 또한 표준정규분포를 따르는 noise 변수를 10개, 0과 1 사이의 균일분포를 따르는 noise 변수를 10개 추가로 만들고 그에 따른 데이터를 랜덤하게 1000개 생성하여 다양한 분석 방법을 통해 주변수인 x_1, x_2가 얼마나 많이 선택되고, noise 변수가 얼마나 적게 선택되는지를 살펴보았다. 시뮬레이션에 사용된 함수는 아래와 같고, 모형은 [그림3.2]와 같다.

- **Saddle Function** : $f(x_1, x_2) = \dfrac{x_1^2 - x_2^2 + 1}{2}$



Saddle Function

[그림3.2] Saddle 모형

3.1절, 3.2절에서와 마찬가지로 원 데이터에서 복원추출을 이용하여 100개의 부스트랩 데이터 셋을 형성하여 랜덤 포레스트를 이용해 어떤 변수가 주변수로 선택되는지 확인하였다. 각각의 모형에 대해 100번의 실험

결과 주변수로 선택된 변수의 선택 비율은 아래 [표3.3]과 같다. [표3.3]에서 보면 변수 선택법으로 랜덤 포레스트를 사용했을 때, 주변수로 사용되었던 x_1과 x_2가 90% 이상의 비율로 선택되고, noise 변수로 불리는 나머지 변수들은 낮은 평균값을 보였다.

특히 기존에 알려진 변수선택법을 이용한 결과와 비교했을 때, 회귀분석, R-square 등의 방법에 비해 트리 계열(트리, Bagging Tree, 랜덤 포레스트 등)을 사용했을 때, 주변수가 훨씬 더 많이 선택되는 것을 볼 수 있다.

변수 / 모형	변수선택법	모형 변수 (%)		noise 변수 (%)
		x_1	x_2	
Saddle	Corr	5	7	2.4
	R-square	53	54	48.05
	Adj-Rsq	29	28	21.95
	CP	14	14	7.8
	Forward	13	11	5.1
	Backward	13	11	5.1
	Stepwise	13	11	5.1
	Tree	100	100	1.75
	Bagging Tree	100	100	12.86
	Bumping Tree	100	100	11.7
	Random Forest	91.1	90.95	21.88

[표3.3] Saddle 모형의 변수 선택 결과

3.4 모형별 결과 비교

프리드만의 모형, 장영재의 모형, Saddle 모형에 대한 변수 선택 결과를 비교해볼 때, Radial, Model 2, Model 2a, Model 3, Saddle Function과 같이 비선형적인 형태를 나타내는 주변수를 포함하는 경우, 기존의 변수 선택 방법에 비해 트리 계열의 방법을 사용했을 때 주변수가 선택되는 비율이 훨씬 높아지는 것을 볼 수 있다.

반면에 Model 0과 같이 각각의 주변수에 가중치를 준 형태의 선형 결합인 경우 회귀분석, 상관분석 등의 변수 선택 방법에서는 모든 주변수가 높은 비율로 선택되지만 트리 계열의 방법에서는 상대적으로 가중치가 낮은 형태의 주변수인 x_1, x_2는 선택 비율이 낮아지는 것을 볼 수 있다.

대부분의 경우에 트리, Bagging Tree, Bumping Tree 방법을 이용하여 선택한 주변수의 비율에 비해 랜덤 포레스트를 이용하여 선택한 주변수의 비율이 낮게 나타나는 것을 볼 수 있다. 트리와 부스트랩 데이터 셋을 활용하는 Bagging Tree의 경우에는 전체 변수를 이용하여 가장 영향력 있는 변수를 선택하여 분할하기 때문에 주변수를 선택하는 빈도가 많아진다. 또한 Bumping Tree의 경우 부스트랩 데이터 셋을 이용하여 분석한 결과 중 가장 좋은 모델을 선택하여 결과에 반영하는 것이므로 마찬가지로 주변수를 선택하는 빈도가 많아진다. 반면에 랜덤 포레스트의 경우에는 부스트랩 데이터 셋을 형성하고, 하나의 데이터 셋을 이용하여 변수를 선택할 때, 1000번 정도의 트리를 생성하고 트리의 노드를 분할할 때에는 기존의 방법과 달리 몇 개의 변수만을 이용하므로 상대적으로 주변수를 선택하는 빈도가 낮게 나타날 수밖에 없게 된다.

위의 분석 결과를 noise 변수 관점에서 보면 랜덤 포레스트는 bumping tree나 bagging tree와 같이 다른 트리계열에 비해 noise를 선택하는 비율이 낮게 나타난다. 이는 랜덤 포레스트의 알고리즘 특성상 수많은 변수들 중에 중요한 주변수를 선택하는 기준에서는 다른 트리계열에 비해 약하지만 수많은 변수들 중에 필요하지 않은 noise 변수를 걸러주는 데에는 다른 트리계열에 비해 강점을 가지고 있다고 할 수 있다.

4. 화학 물질 관련 자료 분석

4.1 데이터 설명

변수 선택에 대한 본 데이터로 화학물질인 Avicel(아비셀)의 용해도 관련 자료를 사용하였다. 이 데이터는 총 47개의 데이터와 Avicel에 포함된 여러 가지 성분 관련 수치로 이루어진 411개의 설명 변수로 구성이 되어 있다. 이 분석에서 사용할 반응 변수는 Avicel 물질의 용해도를 나타내는 수치로 연속형 변수이다.

본 논문에서는 연속형 반응 변수를 가지고 랜덤 포레스트 방법으로 용해도를 가장 잘 설명할 수 있는 변수를 선택하고, 기존의 다른 방법들과 비교해 보았을 때, 얼마나 잘 예측하는지를 살펴보고자 한다. 설명 변수들에 대한 설명의 예시는 아래의 [표4.1]과 같다.

변수명	변수 설명	구분
A6	Average Bonding Information content (order 0)	연속
A7	Average Bonding Information content (order 1)	연속
A8	Average Bonding Information content (order 2)	연속
⋮	⋮	
C1	(1/2)X BETA polarizability (DIP)	연속
C2	(1/6)X GAMMA polarizability (DIP)	연속
C3	1X BETA polarizability (DIP)	연속
⋮	⋮	

[표4.1] Avicel data의 독립 변수

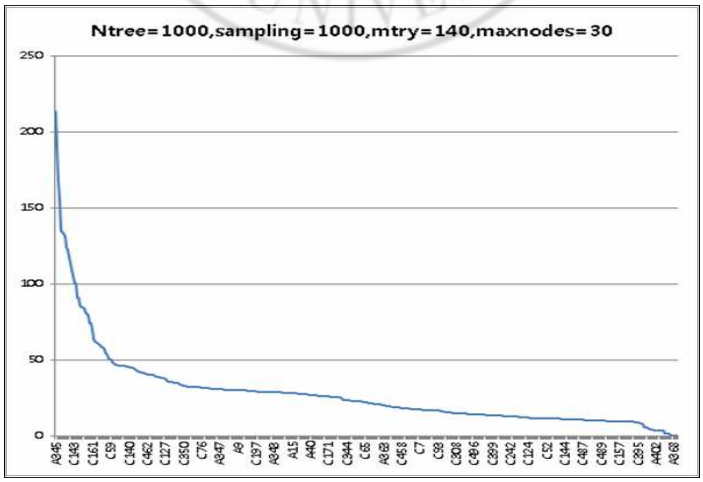
4.2 주변수 선택

본 연구에서는 Avicel 데이터를 의사결정나무의 분리기준을 사용하여 원데이터에서 복원추출을 이용하여 1000개의 부스트랩 데이터 셋을 형성하였고, 랜덤 포레스트의 경우, 여러 가지 옵션을 지정해줄 수 있기 때문에 아래 [표4.2]와 같이 지정한 후, 선택된 변수들을 살펴보았다.

아래 [표4.2]에서 ntree는 랜덤 포레스트 방법에서 사용할 트리의 수를 의미하고, sampling은 1000개의 부스트랩 데이터 셋 중 사용한 데이터 셋의 개수를 의미한다. 또한, mtry는 각각의 노드를 분할할 때, 랜덤으로 선택되는 변수의 개수를 의미하며 maxnodes는 각각의 트리에서 최대로 사용할 terminal node의 개수를 의미한다.

옵션명	ntree	sampling	mtry	maxnodes
값	1000	1000	140	30

[표4.2] 랜덤 포레스트 방법에 사용한 옵션



[그림4.1] 랜덤 포레스트 분석 결과 변수 선택 횟수

위의 [그림4.1]은 1000개의 부스트랩 Avicel 데이터에 대한 랜덤 포레스트 분석 결과 중 각각의 변수가 선택된 횟수를 나타낸 그래프이다. 그래프를 보면 y축의 값이 50일 때, 기울기가 완만해지는 것을 알 수 있고, 이를 이용하여 총 411개의 변수 중 변수 선택횟수가 50이상인 값을 가진 37개의 변수들을 주변수로 선택하여 분석하였다. 주변수로 선택된 변수는 아래 [표4.3]과 같다.

선택 변수명						
A345	C1	C143	C441	C161	C183	C59
A157	C181	C100	C96	C303	C333	
A393	C467	A396	C402	C255	C133	
A357	C168	A355	C138	C169	C505	
A398	C18	C349	C217	C398	C201	
C3	C257	C164	C356	C291	C269	

[표4.3] 랜덤 포레스트 결과 주변수로 선택된 변수

4.3 모형 적합 및 결과 비교

본 연구에서 사용한 랜덤 포레스트 방법은 bagging 방법에서 트리들 간의 상관관계를 줄이면서 그로 인해 생기는 분산을 줄이고자하는 목적에서 연구되었다. 따라서 Avicel 데이터를 bagging 방법과 랜덤 포레스트 방법으로 각각 41개의 주변수를 선택하고, 선택한 변수를 이용하여 회귀분석과 의사결정나무 분석을 실시하였다.

Stepwise를 통한 회귀 분석 결과는 아래 [표4.4]와 [표4.5]에 나타난 것과 같다. [표4.4]와 [표4.5]를 비교해보면 [표4.4]보다는 [표4.5]에서의 결과가 Adj R-Sq의 값은 높게 나타나고, MSE의 값과 Coeff Var의 값은 낮게 나타나는 것을 볼 수 있다.

[표4.6]은 bagging과 랜덤 포레스트를 이용하여 선택된 변수를 이용하여 의사결정나무 분석을 실시한 결과이다. 이 분석에서는 두 개의 결과가 거의 비슷하게 나타나는 것을 볼 수 있다.

선택 변수 : A9, C57, C12, C18, A358, A357					
Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	6	455.28327	75.88055	8.95	<.0001
Error	40	339.26992	8.48175		
Total	46	794.55319			

Root MSE	2.91234	R-Square	0.5730
Dependent Mean	3.65957	Adj R-Sq	0.5090
Coeff Var	79.58150		

[표4.4] Bagging을 통해 선택된 변수를 이용한 회귀분석 결과

선택 변수 : A157, A345, C164, C291, C255, C505, C59					
Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	7	482.65899	68.95128	8.62	<.0001
Error	39	311.89420	7.99729		
Total	46	794.55319			

Root MSE	2.82795	R-Square	0.6075
Dependent Mean	3.65957	Adj R-Sq	0.5370
Coeff Var	77.27531		

[표4.5] 랜덤 포레스트를 통해 선택된 변수를 이용한 회귀분석 결과

Bagging을 이용한 변수		랜덤 포레스트를 이용한 변수	
Statistic	**Value**	**Statistic**	**Value**
AVE SQ ERR	7.8061	AVE SQ ERR	7.9739
ASE/ASE(root)	0.4618	ASE/ASE(root)	0.4717
R-Square	0.5382	R-Square	0.5283

[표4.6] Bagging과 랜덤 포레스트를 통해 선택된 변수를 이용한 의사결정나무 분석 결과

다음으로는 Avicel 자료의 설명변수 411개를 모두 포함하여 bagging과 랜덤포레스트, stepwise 분석을 각각 실시하여 세 개의 분석 결과에서 나온 변수들을 취합하였다. 앞의 3장에서 시뮬레이션 자료를 가지고 분석한 결과, 트리계열의 분석과 회귀분석은 주변수를 뽑는 성향에 차이가 있었다. 이러한 차이를 반영하여 3가지 분석에서 하나라도 선택된 변수는 주변수로 포함하고, 3가지 분석에서 모두 선택되지 않은 변수들은 noise 변수라고 판단하고 변수 선택에서 제외하였다.

선택한 변수들은 아래 [표4.7]과 같고, 전체 411개의 설명변수와 [표4.7]의 설명변수를 이용하여 각각 MARS(Multivariate Adaptive Regression Splines) 분석을 실시하고, 그 결과를 비교하였다.

선택 변수						
A157	A396	C121	C18	C269	C345	C505
A345	A398	C133	C181	C274	C349	C507
A355	A400	C137	C183	C291	C356	C511
A357	A401	C138	C194	C299	C398	C57
A359	A402	C143	C2	C3	C402	C59
A360	A6	C161	C201	C303	C404	C6
A361	A9	C164	C217	C333	C425	C75
A363	C1	C168	C255	C339	C441	C76
A393	C100	C169	C257	C344	C467	C96

[표4.7] MARS 분석에 사용될 변수

분석을 실시한 후의 결과는 아래 [표4.8], [표4.9]와 같다. [표4.8]은 전체 411개의 변수를 이용한 것으로 최종적으로 총 6개의 변수가 선택되었고, GCV 통계량 값은 12, RSS 값은 258, R-Square 값은 0.67이 나온다는 것을 보여준다. 또한 [표4.9]는 [표4.7]의 선택변수 63개를 이용하여 분석을 실시한 결과로서 GCV는 12, RSS 값은 228, R-Square 값은 0.71로 전체 변수를 대상으로 분석을 실시한 결과보다 더 나은 결과를 보였다.

함수식	통계량	값	선택변수
y1 = 7.5 - 0.14 * max(0, 59 - A345) - 0.1 * max(0, A393 - 35) + 269 * max(0, 0.017 - C59) - 15 * max(0, C171 - 104) - 72 * max(0, 104 - C171) + 344 * max(0, C181 - 2.1) - 185 * max(0, C363 - 3.8)	GCV	12	A345 A393 C59 C171 C181 C363
	RSS	258	
	RSq	0.67	

[표4.8] 전체 변수 대상 MARS 분석

함수식	통계량	값	선택변수
y1 = 1.2 - 0.29 * max(0, 45 - A345) - 0.11 * max(0, A393 - 35) + 41 * max(0, C6 - 0.29) - 89 * max(0, 0.29 - C6) + 2547 * max(0, C18 - 6.3e-0.7) + 197 * max(0, 0.026 - C59) + 9.2 * max(0, 0.79 - C121) - 0.021 * max(0, C137 - 3002)	GCV	12	A345 A393 C59 C6 C18 C121 C137
	RSS	228	
	RSq	0.71	

[표4.9] 선택 변수 대상 MARS 분석

[표4.9]에서 선택된 변수에 대한 설명은 아래 [표4.10]과 같다.

변수명	내용
A345	Molecular weight
A393	Relative molecular weight
C59	FHASA Fractional HASA (HASA/TMSA) [Quantum-Chemical PC]
C6	Average Bonding Information content (order 0)
C18	Avg 1-electron react. index for a C atom
C121	HASA-2 [Quantum-Chemical PC]
C137	Highest normal mode vib frequency

[표4.10] MARS 분석 결과 선택된 변수 설명

5. 결론

본 연구에서는 프리드만의 모형, 장영재의 모형, 기타 모형의 형태를 가진 시뮬레이션 자료를 이용하여 여러 가지 분석 방법에 대한 결과 비교를 통해 주변수와 noise 변수의 선택 경향을 파악하였다. 이러한 시뮬레이션 결과를 이용하여 실제 Avicel이라는 화학 물질의 용해도를 측정하기 위한 분석을 하였고, 다음과 같은 방법으로 결과를 비교하였다.

먼저 bagging 방법론에서 시작된 랜덤 포레스트와 bagging과의 비교를 위하여 각각의 방법으로 주변수를 선택하고, 선택된 변수들을 가지고 stepwise와 의사결정나무 분석을 실시하였다. 그 결과, 랜덤 포레스트 방법으로 선택된 변수들에 대한 회귀분석 결과가 R-Square 값은 더 높게, MSE와 Coeff Var 값은 더 낮게 나타났다. 또한 의사결정나무의 분석 결과는 두 가지 방법이 비슷하게 나타났다.

다음으로 bagging, 랜덤 포레스트, stepwise 분석으로 각각의 주변수를 선택하고 각각의 방법으로 선택된 변수들을 취합하여 선택되지 않은 변수들은 noise 변수라 판단하고 분석에서 제외하였다. 총 63개의 변수로 MARS 분석을 실시하였고, 전체 변수를 대상으로 한 MARS 분석 결과와 비교하였을 때, 선택된 변수도 다르고 모형의 성능도 더 좋아졌다.

이를 통해 분석에 필요하지 않은 noise 변수를 잘 선택하여 제거하고, 분석에 필요한 주변수를 잘 선택하는 것이 모형의 성능을 결정하는 중요한 요인이라는 것을 알 수 있고, 분석에 필요하지 않은 noise 변수를 찾는 데에는 랜덤 포레스트 방법이 기존의 다른 방법들에 비해 더 효과적이라는 것을 알 수 있다.

참고문헌

[1] Andy Liaw and Matthew Wiener (2002) 'Classification and Regression by randomForest', R News, Vol.2/3, December 2002.

[2] Trevor Hastie · Robert Tibshirani · Jerome Friedman, The Elements of Statistical Learning, Springer, p.217~220, 246~247, 587~604.

[3] 이상원 (2011), 'bumping을 이용한 변수 선택'.

[4] 천희경 (2009), 'bagging을 이용한 변수선택 방법'.

[5] 김동일 (2007), '데이터 마이닝에서 변수 선택 방법 비교 =Comparison of variable selection methods'.

[6] 정성석 · 김순영 · 임한필, '의사결정나무에서 분리 변수 선택에 관한 연구', 응용통계연구 제 17권 2호, 2004년, p.347~357.

[7] 이영섭 · 오현정 · 김미경, '데이터 마이닝에서 배깅, 부스팅, SVM 분류 알고리즘 비교 분석', 응용통계연구 제 18권 2호, 2005년, p.343~354.

[8] 박창이 · 김용대 · 김진석 · 송종우 · 최호식, 'R을 이용한 데이터마이닝', 교우사, p.259~290.