The Korean Journal of Applied Statistics
2021, Vol. 34, No. 2, 153–166

DOI: https://doi.org/10.5351/KJAS.2021.34.2.153
Print ISSN1225-066X / Online ISSN 2383-5818

Feature selection and prediction modeling of drug responsiveness in Pharmacogenomics

Kyuhwan Kim[a], Wonkuk Kim[1,b]

[a]Department of Statistics, Chung-Ang University; [b]Department of Applied Statistics, Chung-Ang University

Abstract

A main goal of pharmacogenomics studies is to predict individual's drug responsiveness based on high dimensional genetic variables. Due to a large number of variables, feature selection is required in order to reduce the number of variables. The selected features are used to construct a predictive model using machine learning algorithms. In the present study, we applied several hybrid feature selection methods such as combinations of logistic regression, ReliefF, TurF, random forest, and LASSO to a next generation sequencing data set of 400 epilepsy patients. We then applied the selected features to machine learning methods including random forest, gradient boosting, and support vector machine as well as a stacking ensemble method. Our results showed that the stacking model with a hybrid feature selection of random forest and ReliefF performs better than with other combinations of approaches. Based on a 5-fold cross validation partition, the mean test accuracy value of the best model was 0.727 and the mean test AUC value of the best model was 0.761. It also appeared that the stacking models outperform than single machine learning predictive models when using the same selected features.

Keywords: AUC, ensemble method, machine learning, random forest, stacking

1. 서론

인간 게놈 프로젝트가 완료되고 DNA 서열 분석(DNA sequencing)기술이 발달함에 따라, 약물유전체학(Pharmacogenomics) 데이터를 기반으로 한 분석이 등장하기 시작했다. 약물유전체학 데이터를 이용한 질병 혹은 약물 반응성 예측에 대한 접근법으로는 대표적으로 다원유전적 위험 평가(Polygenic risk scoring)와 머신러닝(machine learning)이 있는데, 다원유전적 위험 평가는 예측성능에 있어 한계를 보이며 (Ho, Daniel Sik Wai 등, 2019), 이에 따라 높은 예측정확도를 보여주는 머신러닝 기반의 다양한 예측모형이 등장하고 있다. 이러한 약물유전체 자료와 임상자료로 구축되는 예측모형은 개인 맞춤형 의료(personalized medicine) 혹은 정밀 의료(precision medicine)와 같은 미래 의학의 일부분으로 여겨지며 현재 활발한 연구가 진행되고 있다.

보통 약물유전체학 데이터는 작은 표본 수에 비해 단일염기 다형성(single nucleotide polymorphism; SNP)의 수가 매우 많아 머신러닝 기반의 예측모형은 변수선택 과정을 필수적으로 수반한다. 변수선택 방법은 크게 필터(filter) 방법, 래퍼(wrapper) 방법, 임베디드(embedded) 방법, 그리고 혼합(hybrid) 방법이 있다. 필터

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2019.
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant no. NRF-2018R1D1A1B07050012).
[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-Ro, Dongjak-Gu, Seoul 06974, Korea. E-mail: wkim@cau.ac.kr

Published 30 April 2021 / journal homepage: http://kjas.or.kr

방법은 모델 학습과는 독립적인 변수선택 방법으로 다른 방법보다 계산이 효율적이기 때문에 (Li 등, 2017) 약물유전체학 데이터의 변수선택에서 가장 많이 이용되며 (Ho, Daniel Sik Wai 등, 2019), 카이제곱 검정이나 코크란-아미티지 검정(cochran-armtiage test), 로지스틱 회귀분석과 같은 연관성 검정의 p 값(p value)을 사용하는 변수 선택 방법과 k-최근접이웃 알고리즘(k-nearest neighbors; KNN) 방법을 이용하는 Relief기반 알고리즘 (Kira과 Rendell, 1992) 등을 포함한다. 래퍼 방법은 예측모형에 쓰이는 변수의 여러 부분집합을 탐색하여 가장 좋은 성능을 보이는 변수의 부분집합을 선택하는 변수 선택 방법으로 높은 성능을 보이지만 계산 비용이 매우 크다. 전진 선택법(forward selection), 후진 제거법(backward elimination), 단계별 선택법(stepwise selection) 등의 방법이 이에 해당되고 약물유전체 자료와 같은 고차원 자료에 응용은 한계가 있다. 임베디드 방법은 모형의 훈련 과정 안에서 예측의 정확도에 기여하는 변수를 찾아내는 변수 선택 방법으로 Tibshirani (1996)가 제안한 라쏘(least absolute shrinkage and selection operator; LASSO)와 같은 L1-정규화(L1-regularization) 방법이 널리 이용되고 있다. Zou 등 (2005)은 LASSO를 일반화하여 L2-정규화(L2-regularization) 항을 포함한 신축망(elastic net) 모형을 제안하였고 이 모형도 변수 선택에 이용될 수 있다. 의사결정나무와 Breiman (2001)이 제안한 랜덤포레스트(random forest)의 결과물인 변수 중요도 측도(variable importance measure)를 기준으로 변수를 선택할 수 있으며 이 방법도 임베디드 방법중의 하나이다. 혼합 방법과 같이 여러 가지의 변수 선택 방법을 혼합하여 변수 선택을 진행할 수도 있다. 한 예로 Wei 등 (2013)은 먼저 필터 방법인 로지스틱 회귀분석으로 얻은 SNP의 p 값을 이용하여 변수선택을 한 후, 남은 변수들에 임베디드 방법인 L1-정규화를 적용하여 추가적인 변수 선택을 진행하였다. Mieth 등 (2016)은 임베디드 방법인 서포트벡터머신(support vector machine; SVM)을 이용하여 SNP를 선택한 후, 필터 방법인 카이제곱 검정으로 얻은 해당 SNP의 p 값에 따라 추가적인 변수 선택을 진행하였다. Jović 등 (2015)은 다양한 분야의 데이터에 대한 변수 선택 방법의 성능을 논의하였으며 Muštra 등 (2012)은 SNP 데이터의 경우 카이제곱 방법과 ReliefF 등이 좋은 성능을 보이는 것을 발표하였다.

선택된 SNP에 적용되는 모형 또한 서포트벡터머신 (Wei 등, 2013; Abraham 등, 2014; Nguyen 등, 2015; Mieth 등, 2016), 랜덤포레스트 (López 등, 2017)와 같은 머신러닝 모형부터, 인공신경망(neural network)과 같은 딥러닝(deep learning) 모형 (Montañez 등, 2018; Romagnoni 등, 2019) 등으로 다양하게 사용할 수 있다. Wei 등 (2013)은 크론병 예측을 위해 앞서 선택한 SNP에 서포트벡터머신을 적용하였고 이 모델은 0.86의 AUC를 보였다. Mieth 등 (2016)은 T2D 예측에 서포트벡터머신을 적용하였고 이 모델은 0.84의 AUC를 보였다. 서포트벡터머신의 핵(kernel)으로는 선형핵(linear kernel)이 주로 이용되었으며 (Wei등, 2013; Nguyen 등, 2015; Mieth 등, 2016), Abraham 등 (2014)은 L1-벌점(L1-penalized) 서포트벡터머신을 이용하여 셀리악병 (Celiac disease)을 예측하여 0.87의 AUC를 보이기도 했다. López 등 (2017)은 랜덤포레스트를 이용하여 T2D를 예측하여 0.85의 AUC를 달성하였는데, 랜덤포레스트는 약물유전체학 연구에서 다른 머신러닝 모형들에 비해 뛰어난 성능을 보여주는 것으로 알려져 있다 (Chen과 Ishwaran, 2012; Boulesteix 등, 2012; Austin 등, 2013; López 등, 2017). (Montañez 등 (2018)은 비만 예측을 위해 인공신경망을 적용하였고 0.99의 AUC를 달성하기도 했다.

위에 설명한 다양한 개별 머신러닝 방법으로 예측모형을 만들 수 있으나 이 논문에선 다양한 머신러닝의 예측모형을 앙상블하는 방법을 적용한다. Wolpert (1992)는 level-0와 level-1 자료의 개념을 통해 여러 개의 다양한 형태의 분류모형을 앙상블(ensemble)하는 방법인 스태킹(stacking)을 제안하였고 과적합(overfitting)을 방지하기 위하여 k-폴드 교차 검증 자료를 사용하여 메타모형(meta learner)를 훈련하는 방법을 제안하였다. 후에 Breiman (1996)은 스태킹 방법을 회귀문제로 확장하였다. 그 후 van der Laan 등 (2007)은 이론적으로 k-폴드 교차 검증 자료로 훈련하여 얻어진 스태킹 모형의 성능이 점근적으로 각각의 모형의 성능보다 높은 것을 증명하였다.

이 논문은 뇌전증 약물유전체 자료를 사용하여 4가지의 다른 절차에 따른 변수 선택을 진행하여 환자

의 약물반응에 대한 3가지 단일 예측모형과 스태킹 모형을 구축하여 총 16가지 모형의 성능을 비교한다. 이 논문의 구성은 다음과 같다. 2장에서는 이 논문에서 사용하는 변수선택 방법인 로지스틱 회귀분석, Relief, 랜덤포레스트를 간략하게 설명하고 다양한 분류기(classifier)를 앙상블하는 스태킹 방법에 대하여 설명한다. 3장에서는 뇌전증 약물유전체 자료와 분석 절차에 대하여 설명하고, 4장에서는 수행한 결과를 설명한다. 마지막 5장에서는 이 논문에 대한 결론을 제시한다.

2. 방법론

2.1. 로지스틱 회귀분석

변수 선택 방법 중 필터 방법의 몇 가지는 로지스틱 회귀분석 모형으로 표현될 수 있다. 최종 변수 선택에 강제로 포함하는 변수 벡터를 $\mathbf{z}$라고 하였을 때, 각 변수 x_j의 랭크는 다음과 같은 로지스틱 회귀분석 모형으로부터 β_j의 p-값으로 계산할 수 있다.

$$\log \frac{P(Y=1|x_j,\mathbf{z})}{1-P(Y=1|x_j,\mathbf{z})} = \beta_0 + \beta_j x_j + \gamma^T \mathbf{z}.$$

변수 x_j가 범주형인 경우 가변수(dummy variable) 열벡터이고 β_j는 행벡터이다. 약물유전체학에서 자주 사용되는 코크란-아미티지 검정의 경우 변수 x_j의 코딩에 따라 다양한 유전모형에 대한 분석을 진행할 수 있다. 예로 변수 x_j의 값이 유전자형(genotype) AA, AB, BB가 가능한 경우, 변수 x_j의 코딩을 0, 1, 2로 하면 부가(additive) 모형이 되고 코딩을 0, 1, 1로 하면 우성(dominant) 모형이 된다. 만약 변수 x_j를 범주형 변수로 간주하면 카이제곱 검정의 p-값을 사용하는 것과 일치하게 된다.

2.2. Relief 기반 알고리즘

Relief 기반 알고리즘은 상호작용을 가지는 변수들을 찾아내는 최근접이웃 알고리즘 기반의 변수 선택 알고리즘이다. 이 알고리즘은 Kira과 Rendell (1992)에 의해 처음으로 제안되었다. 훈련용 데이터가 $\mathcal{X} = \{\mathbf{x}_i \in \mathbf{R}^d, y_i \in \mathbf{R} : 1 \le i \le n\}$이고 반응 변수는 $y_i = 0$ 또는 1로 두 클래스 분류 문제인 경우, Relief 알고리즘은 다음과 같은 과정을 반복한다.

1. 최초에 d 차원의 변수 가중치 벡터(weight vector)를 $\mathbf{w}^{(0)} = (w_1, \ldots, w_s, \ldots, w_d) = (0, \ldots, 0)$와 같은 영벡터로 설정한다.
2. 임의표본 $(\mathbf{x}_i, y_i)$을 추출하고, 이 임의표본의 같은 클래스의(같은 y_i 값) 최근접이웃인 최근접성공(nearest hit) 표본 $\mathbf{x}_{i_h} = (x_{i_h 1}, \ldots, x_{i_h d})$와 다른 클래스의(다른 y_i 값) 최근접이웃인 최근접실패(nearest miss) 표본 $\mathbf{x}_{i_m} = (x_{i_m 1}, \ldots, x_{i_m d})$를 구한다. 만약 x_j가 연속형 변수이며 정규화(normalization) 되어 있다면, 두 표본 사이의 거리는 다음과 같다.

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{d} |x_{1j} - x_{2j}|.$$

3. $(r+1)$번째 반복에서 가중치 벡터를 다음과 같이 갱신한다.

$$w_j^{(r+1)} = w_j^{(r)} - \frac{|x_{ij} - x_{i_h j}|}{R} + \frac{|x_{ij} - x_{i_m j}|}{R}, \quad j = 1, \ldots, d.$$

이 때, R은 총 반복 횟수이다.

ReliefF 알고리즘 (Kononenko, 1994)는 Relief 알고리즘을 개선한 것이다. Relief 알고리즘이 각 반복에서 한 개의 최근접성공과 한 개의 최근접실패를 사용한 것과는 다르게, ReliefF는 k개의 최근접성공과 k개의 최근접실패를 사용한다. 이 때문에 ReliefF 알고리즘은 Relief 알고리즘보다 로버스트하다. 보통의 경우 ReliefF 알고리즘은 10개씩의 최근접성공과 최근접실패를 사용한다. ReliefF 알고리즘은 가중치 벡터를 갱신하는 방법 외엔 Relief 알고리즘과 거의 유사하다. $H_i = \{\mathbf{x}_{i_{h1}}, \ldots, \mathbf{x}_{i_{hk}}\}$를 $\mathbf{x}_i$의 k개의 최근접성공의 정규화 된 집합이라 하고, $M_i = \{\mathbf{x}_{i_{m1}}, \ldots, \mathbf{x}_{i_{mk}}\}$를 $\mathbf{x}_i$의 k개의 최근접실패의 정규화 된 집합이라고 할 때, $(r+1)$번째 반복에서 가중치 벡터는 다음과 같이 갱신한다.

$$w_j^{(r+1)} = w_j^{(r)} - \frac{1}{k}\sum_{\mathbf{x}_{i'} \in H_i} \frac{|x_{ij} - x_{i'j}|}{R} + \frac{1}{k}\sum_{\mathbf{x}_{i'} \in M_i} \frac{|x_{ij} - x_{i'j}|}{R}, \quad j = 1, \ldots, d.$$

그러나, ReliefF 알고리즘의 성능은 잡음이 있는 변수들에 의해 저하되는 것으로 알려져 있다. 본 논문에서도 사용된 Tuned ReliefF 알고리즘 (Moore과 White, 2007)은 ReliefF 알고리즘을 여러 번 반복하며 낮은 가중치를 보이는 잡음이 있는 변수를 제거함으로써 ReliefF 알고리즘을 개선한 것이다. Tuned ReliefF 알고리즘의 과정은 다음과 같다:

1. ReliefF 알고리즘을 적용한 후(보통, $k = 10$), 가중치 w에 따라 내림차순으로 정렬한다.
2. 가장 가중치가 낮은 변수를 제거한다.

만약 반복마다 10%의 가장 성능이 낮은 변수를 제거하면 이를 TuRF 10%라고 한다. 변수의 수가 적은 다중차원축소(multifactor dimensionality reduction; MDR) 분석의 경우, TuRF 알고리즘이 ReliefF 알고리즘보다 더 좋은 성능을 보여주는 것으로 알려져 있다 (Moore과 White, 2007).

2.3. 랜덤포레스트

의사결정나무 모형은 비선형 모형의 하나로 이해하기 쉬운 높은 해석력을 보이지만 하나의 의사결정나무 모형은 예측력이 높지 않아 예측모형으로써 활용성은 떨어진다. Breiman (1996)은 부스트랩(bootstrap) 자료를 활용하여 많은 의사결정나무 모형을 만든 후 평균 혹은 다수결을 기반으로 하는 배깅(bagging)이라는 분석방법을 제안하였고 후에 의사결정나무 모형들 간의 상관관계를 낮추어 평균제곱오차(mean square error)를 작게 할 수 있는 방법인 랜덤포레스트 (Breiman, 2001)를 제안하였다. 랜덤포레스트 에서는 노드(node)에서 이분할 하는 변수를 랜덤하게 추출하여 선택된 변수만을 사용하여 각각의 의사결정나무 모형을 생성한다. 각각의 노드에서 분할 시 새로운 변수를 선택하여 생성된 의사결정나무 모형들 간의 상관관계를 낮게 해주는 효과를 가져와 보통 랜덤포레스트가 모든 변수를 사용하여 노드에서 이분할 하는 배깅 모형보다 예측력이 뛰어나다. 랜덤포레스트는 하나의 의사결정나무 모형보다 해석력이 낮아지는 단점이 있지만 변수 중요도 측도를 사용하여 각 변수의 중요도를 측정할 수 있으며 이는 변수 선택에 활용할 수 있다. 널리 쓰이는 변수 중요도 측도로는 평균불순도감소(mean decrease impurity)와 순열변수중요도(permutation importance)가 있다.

변수 x_j의 평균불순도감소는 각각의 의사결정나무에서 변수 x_j가 분할에 사용된 노드에서 감소된 불순도 측도의 평균으로 정의된다. 의사결정나무 모형의 각 노드 t에서 불순도 측도를 $i(t)$, 변수 x_j에 의해 분할되기 이전의 노드에 포함된 표본의 개수를 n_t, 분할 s_t후의 노드를 각각 t_L과 t_R, 분할 이후 각 노드에 속한 표본의 개수를 n_{t_L}과 n_{t_R}이라고 하였을 때, 변수 x_j의 불순도 측도 감소분 $\Delta i(s,t)$는 다음과 같다.

$$\Delta i(s,t) = i(t) - \frac{n_{t_L}}{n_t} i(t_L) - \frac{n_{t_R}}{n_t} i(t_R).$$

각 의사결정나무 모형을 T라 할 때, n_T개의 의사결정나무를 사용한 랜덤포레스트에서 x_j의 평균불순도 감소

$\mathrm{VI}_1(x_j)$은 다음과 같이 정의된다.

$$\mathrm{VI}_1(x_j) = \frac{1}{n_T} \sum_T \sum_{t \in T: v(s_t)=x_j} \frac{n_t}{n} \Delta i(s,t).$$

이 때, n은 전체 표본 개수이고, $v(s_t)$는 분할 s_t에서 사용된 변수이다. 불순도 측도를 정확도(accuracy)로 할 경우 이 때의 변수 중요도를 평균정확도감소(mean decrease accuracy)라고 하며 불순도 측도를 지니 계수(gini index)로 할 경우 이 때의 변수중요도를 평균지니감소(mean decrease gini)라고 한다. 분류 문제의 경우 이 두 가지의 변수 중요도 측도가 널리 사용된다. 변수 x_j의 순열 변수 중요도는 랜덤포레스트의 각 의사결정나무 모형에서 변수가 무작위로 분포될 때 저하되는 예측력의 평균으로 정의된다. 저하되는 예측력은 (out of bag; OOB) 표본을 이용하여 계산한다. 각 의사결정나무 모형에서 변수 x_j가 무작위로 분포될 때의 예측력 감소분 $\Delta P(x_j)$ 다음과 같이 계산한다.

1. 의사결정나무 모형의 예측력(정확도, 지니 계수 등) P를 OOB 표본을 이용하여 계산한다.
2. OOB 표본의 j 번째 변수의 데이터를 무작위로 섞은 후 이 데이터를 이용하여 의사결정나무 모형의 예측력 P_j를 계산하고 예측력의 감소분 $\Delta P(x_j) = P - P_j$를 구한다.

각 의사결정나무 모형을 T라 할 때, N_T개의 의사결정나무를 사용한 랜덤포레스트에서 x_j의 순열변수중요도는 다음과 같이 정의한다.

$$\mathrm{VI}_2(x_j) = \frac{1}{N_T} \sum_T \Delta P^{(T)}(x_j).$$

이 때, $\Delta P^{(T)}(x_j)$는 나무 T에서의 예측력 감소분이다. 이러한 변수 중요도의 크기에 따라 변수를 선택할 수 있으며 VSURF (Genuer 등, 2015)와 varSelRF (Díaz-Uriarte, 2007)와 같은 랜덤포레스트를 이용하는 다양한 변수선택 방법이 제안되었다.

2.4. 스태킹(stacking)

스태킹은 여러 가지의 모형을 결합하는 앙상블 방법의 하나이다. 스태킹이 제안되기 전에는 여러 개의 모형을 결합할 때 과적합 문제와 모형 간의 상관 문제가 발생하여 연구자들은 여러 개의 모형 중 가장 좋은 모형을 선택하는 방법에 집중하였다. 하지만 Wolpert (1992)는 레벨-1(level-1) 데이터의 개념과 교차 검증 방법을 이용하여 분류문제의 스태킹 방법을 제안하였으며, Breiman (1996)은 스태킹 방법을 회귀문제로 확장시켰다. 그 후 van der Laan 등 (2007)은 스태킹 방법을 수학적으로 설명하였고 스태킹으로 결합된 모형이 결합되지 않은 각각의 단일 모형보다 더 높은 성능을 보인다는 것을 점근적으로 증명하였다.

훈련용 데이터를 $\mathcal{X} = \{\mathbf{x}_i \in \mathbf{R}^d, y_i \in \mathbf{R} : 1 \le i \le n\}$로 표현하고 레벨-0(level-0) 데이터라고 부르며 $\mathbf{x}_i$는 설명 변수의 벡터이다. 이 때, M개의 모형을 결합하는 스태킹 과정은 다음과 같다.

1. 레벨-0 데이터를 $\bigcup_{k=1}^{K} \mathcal{X}_k = \mathcal{X}$ 그리고 $\mathcal{X}_k \cap \mathcal{X}_{k'} = \emptyset, \forall k \neq k'$를 만족시키는 K개의 폴드 $\mathcal{X}_1, \ldots, \mathcal{X}_K$로 최대한 크기가 같도록 분할한다.
2. $\mathcal{X}^{(-k)} = \mathcal{X} - \mathcal{X}_k$를 이용하여 m번째 기저 모형(base learner) $f_m^{(-k)}(\mathbf{x})$를 적합하고 다음과 같은 예측치를 만든다.

$$z_{im} = \widehat{f_m^{(-k)}}(\mathbf{x}_i), \quad (\mathbf{x}_i, y_i) \in \mathcal{X}_k, \quad m = 1, \ldots, M.$$

여기서 구한 $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iM})$을 레벨-1 데이터라고 정의한다.

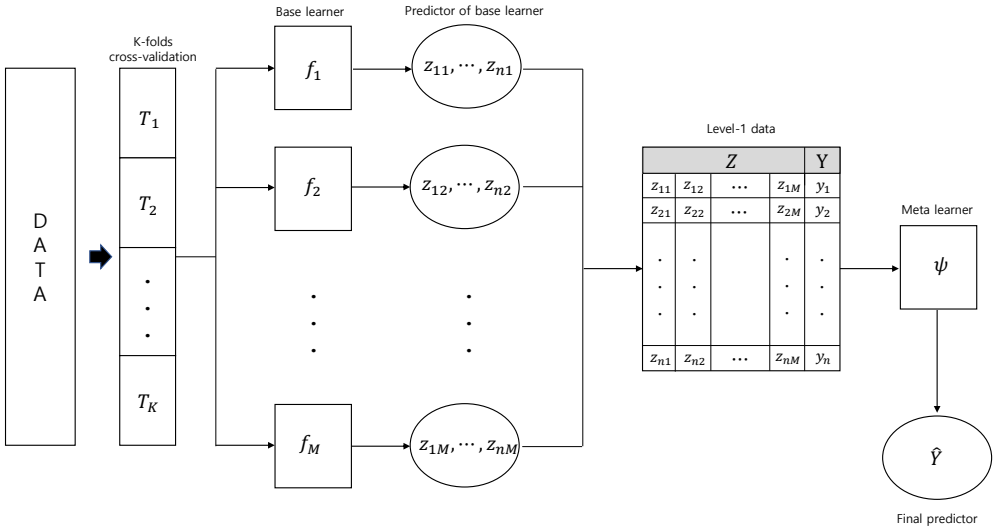


Figure 1: *Flowchart of stacking modeling process.*

3. 위 과정에서 생성된 level-1 데이터를 이용하여 손실 함수 $L(\cdot,\cdot)$를 최소화하는 메타 모형(meta learner) $\hat{g}$을 생성한다.

4. 모든 훈련용 데이터를 사용하여 각 기저 모형을 적합하여 $\hat{f}_m$을 구하고 전 과정에서 생성된 메타 모형으로 결합하여 스태킹 모형 $\hat{g}\left(\hat{f}_1,\ldots,\hat{f}_M\right)$을 예측에 사용한다. 따라서, $\mathbf{x}_0$에서의 예측값은 $\hat{g}\left(\hat{f}_1(\mathbf{x}_0),\ldots,\hat{f}_M(\mathbf{x}_0)\right)$으로 표현할 수 있다.

분류 문제의 경우, 보통 각 클래스에 대한 사후 확률 예측치를 레벨-1 데이터로 사용한다.

3. 데이터 분석

3.1. 데이터 설명

이 논문에서는 뇌전증 환자의 약물반응 예측모형을 구축하기 위하여 최근 발간된 Kang 등 (2019)과 Lee 등 (2020)의 논문에 이용된 400명의 진유전체(exome)자료를 이용하였다. 자료에 대한 자세한 설명은 Kang 등 (2019)에 서술되어 있으며 이 절에서는 간략한 설명을 포함하였다.

대한민국내 3차 뇌전증 의료센터 10곳을 대상으로 뇌전증 환자 400명의 차세대 염기 서열분석을 통한 환자들의 고차원 진유전체의 유전자형 자료가 생성되어 CODA (http://coda.nih.go.kr)에 등록되어 있다. 모든 뇌전증환자는 뇌전증의 임상적 정의 (Fisher 등, 2014)에 의해 정의되었으며, 2년 동안 의료센터의 뇌전증 전문가에 의해 관리되어 왔다. 이 논문에서 반응변수로 사용하는 약물반응은 Kim 등 (2011)이 기술한 정의와 기준에 따라 약물 내성군(drug-resistance group; DR 그룹) 또는 약물 반응군(drug-responsive group; DS 그룹)으로 분류되었다. 약물저항성은 모집하기 이전 1년동안 두 가지 이상의 적절한 최대 허용 용량의 항뇌전증 약물(antiepileptic drugs; AEDs)을 복용하면서 최소 12번의 이유 없는 발작이 발생하는 것으로 정의하였는데, 이는 두 그룹 간의 대조를 강화하기 위해 전통적인 방법보다 더욱 엄격하게 정의한 것이다. 뇌전증 치료를

받은 DR 그룹의 환자들은 수술 결과에 관계없이 DR 그룹으로 분류되었다. 1차 혹은 2차 혈족에서 뇌전증의 발생이 분명한 환자, AED 치료가 미흡한 환자, 운동성 발작이 없는 환자, 의식장애 없이 비운동성 발작만 발생한 환자, 혹은 진행성 DEEs를 가진 환자는 본 연구에서 제외되었다. 약물 반응성은 후속 방문일로부터 적어도 1년 이상 발작에서 자유로워지는 것으로 정의되었다. Lee 등 (2020)의 논문에서와 같이 환자의 나이(age)와 뇌전증 지속기간(duration) 두 변수는 강제적으로 모든 변수선택 과정 및 모형에 포함하였다.

3.2. 실험설계 및 데이터 전처리

유전자자료의 품질관리를 위하여 차세대염기서열 실험결과에서 얻어지는 읽기깊이(read depth; DP)의 값이 10이상인 이대립인자성(biallelic) SNP만 추출하였고 그 개수는 54,708개이다. 또, 본 논문에서는 각 SNP변수를 유전자형에 따라 0, 1, 2로 코딩하였다. 약물유전체학 데이터와 같이 표본 수는 적고 변수는 많은 고차원 데이터의 경우 과적합이 일어날 수 있다. 이러한 문제점을 보완하기 위해, 이 논문에서는 데이터를 5개의 폴드로 나누고 4개의 폴드를 훈련용 데이터로 남은 한 개의 폴드를 검증용 데이터로 사용하였다. 총 표본의 수가 400으로 많지 않은 이유로 각각의 폴드가 모두 한 번씩 테스트 데이터로 쓰이도록 이 과정을 5번 반복하였다. 따라서 320개의 표본이 훈련용 데이터로 80개의 표본이 검증용 데이터로 사용해 다섯 번 반복하여 모형을 구축하였고 이를 통해 검증 정확도를 계산하였다. 고정된 검증용 데이터에서 모든 변수선택은 각각 훈련용 데이터에서 실행하였고 스태킹 모형을 위한 10-폴드 교차검증 자료는 훈련용 데이터를 10개의 폴드로 나누어서 적합하여 각 검증용 데이터는 모형의 평가에만 사용하였다.

3.3. 변수 선택과 예측모형 구축

이 논문에서는 다음과 같은 16가지의 방법으로 변수선택과 예측모형을 구축하여 결과를 비교하였다. 첫번째 변수 선택 방법 FS1에서는 총 54,710개의 변수를 대상으로 임베디드 방법인 랜덤포레스트의 평균정확도감소 값이 높은 대략 10%에 해당하는 5,000개의 변수를 선택한 후 필터 방법인 ReliefF를 적용하여 평균+3표준편차를 넘는 변수를 선택하였다. 대략 훈련용 데이터의 표본수의 약 1/3인 100개의 변수가 선택되었다. 두 번째 변수 선택 방법 FS2의 경우는 ReliefF를 적용하지 않고 랜덤포레스트만으로 100개의 변수를 선택하였다. 세 번째 방법 FS3에서는 각각의 SNP 변수와 나이, 뇌전증 지속기간을 변수를 대상으로 필터 방법인 로지스틱 회귀분석을 적용한 후 p-값이 0.005 이하인 변수와 전체 데이터에 필터방법인 TuRF를 적용한 후 TuRF 점수가 평균+9표준편차를 넘는 변수를 선택한 후, 선택된 변수에 임베디드 방법인 LASSO를 적용시켜 추가적인 변수 선택을 진행하였다. 비교연구를 위하여 기존에 약물유전체학 데이터의 예측모형에 많이 사용되는 방법인 Wei 등 (2013)의 변수선택 방법 FS4를 이용하였다. 이 경우 로지스틱 회귀분석의 p-값이 0.005이하인 변수를 추출하고 선택된 변수에 LASSO 방법을 적용하였다. Wei 등 (2013)은 17만여개의 SNP가 포함된 데이터를 이용하였고 로지스틱 회귀분석을 통한 변수선택 과정에서 p-값의 절사값으로 0.0001을 사용하였으나, 본 연구에서 사용된 훈련용 데이터의 SNP수는 5만여개로 Wei 등 (2013)의 데이터보다 변수의 개수가 적기 때문에 로지스틱 회귀분석을 통한 변수선택 과정에서 p-값의 절사값으로 0.005를 사용하였다. 설명한 4가지 방법으로 선택된 변수를 사용하여 각 예측모형은 랜덤포레스트, 그래디언트 부스팅, 방사형핵 서포트벡터머신, k-최근접이웃 알고리즘, 로지스틱 회귀, 신축망(elastic net)을 적합한 후 스태킹 앙상블을 적용하여 4개의 다른 스태킹 모형 M1E, M2E, M3E, M4E를 생성하였다. 각 변수 선택 방법으로 선택된 변수를 사용하여 3가지의 머신러닝 알고리즘인 랜덤포레스트, 방사형 핵 서포트벡터머신, 그래디언트 부스팅을 적용한 12가지의 예측모형도 생성하였다. 그 중 모형 M4S는 Wei 등 (2013)과 유사한 예측모형으로 볼 수 있다. 변수 선택 방법 FS3와 FS4의 경우 계산시간은 대략 각 방법당 1시간 정도 소요되었고, 랜덤포레스트를 사용하는 FS1과 FS2의 경우 튜닝을 하지 않으면 FS1 또는 FS2와 비슷한 시간이 소요된다. 본 논문에서는 10폴드 교차검증으로 튜

Table 1: The accuracy values and the average accuracy values

Name	Feature Selection	Prediction Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SE
M1E		Stacking	0.725	0.662	0.738	0.709	0.802	0.727	0.023
M1R	FS1: Hybrid	RF	0.700	0.638	0.700	0.671	0.778	0.697	0.023
M1S	(RF, ReliefF)	R–SVM	0.663	0.525	0.613	0.620	0.704	0.625	0.030
M1G		GBM	0.725	0.625	0.786	0.595	0.802	0.707	0.042
M2E		Stacking	0.650	0.600	0.738	0.683	0.556	0.645	0.032
M2R	FS2: Embedded	RF	0.638	0.588	0.738	0.595	0.543	0.620	0.032
M2S	(RF)	R–SVM	0.563	0.538	0.688	0.620	0.580	0.598	0.026
M2G		GBM	0.638	0.588	0.763	0.557	0.642	0.637	0.035
M3E		Stacking	0.633	0.663	0.600	0.675	0.650	0.645	0.013
M3R	FS3: Hybrid	RF	0.550	0.688	0.588	0.570	0.667	0.612	0.027
M3S	(LR, TuRF, Enet)	R–SVM	0.625	0.638	0.575	0.633	0.593	0.613	0.012
M3G		GBM	0.675	0.575	0.688	0.620	0.654	0.642	0.020
M4E		Stacking	0.638	0.663	0.675	0.650	0.638	0.653	0.007
M4R	FS4: Hybrid	RF	0.575	0.625	0.588	0.633	0.667	0.617	0.016
M4S	(LR, LASSO)	R–SVM	0.613	0.600	0.613	0.625	0.686	0.628	0.016
M4G		GBM	0.638	0.625	0.713	0.613	0.725	0.663	0.023

RF: random forest, GBM: gradient boost model, R-SVM: radial kernel support vector machine, knn: k-nearest neighbors, LR: logistic regression, Enet: elastic net

닝을 실행하였고 대략 12시간 정도의 시간이 소요되었다. 계산시간은 3.0 GHz Intel Xeon 10코어 프로세서의 컴퓨터에서 실행한 결과이다.

4. 분석결과

모형의 성능을 비교하기 위해 각 모형의 정확도와 AUC값을 계산하였으며 그 결과는 Table 1과 Table 2에 정리하였다. Figure 2에서는 4개의 스태킹 모형의 평균 ROC곡선을 확인할 수 있고 Figure 3에서는 4개의 스태킹 모형의 각 폴드에서 ROC곡선을 확인할 수 있다. 랜덤포레스트의 변수중요도와 ReliefF의 조합으로 변수선택을 진행하여 여러 머신러닝기저 모형을 스태킹으로 생성한 M1E모형이 평균 정확도가 0.727이고 평균 AUC가 0.761로 다른 모형과 비교하였을 때 높은 성능 보여주었다. 전반적으로 FS1으로 변수선택을 진행한 경우의 정확도와 AUC가 높은 것을 볼 수 있다. 다른 변수 선택 방법의 경우 Wei 등 (2013)이 사용한 변수선택 방법과 스태킹을 적용한 M4E가 AUC = 0.727로 높은 성능을 보여주고 있다. 변수 선택 방법과 관계없이 서포트벡터머신을 사용한 모형의 성능은 전반적으로 낮은 것을 볼 수 있고, 오히려 그래디언트 부스팅으로 학습한 모형의 성능이 상대적으로 높았다. 통계적 유의미성을 확인하기 위하여 AUC값의 로짓(logit) 변환 후 완전임의배치법(randomized complete block design)의 분산분석(ANOVA)을 실행한 결과 변수 선택 방법 요인의 F 통계량은 8.260이고 p-값은 9.03×10^{-5}이었고 스태킹 또는 머신러닝 예측 모형 요인의 F 통계량은 6.970이고 p-값은 3.65×10^{-4}로 통계적으로 유의미한 차이를 보여주고있다.

Tukey의 HSD를 이용한 다중비교 결과 FS1의 변수선택방법은 다른 세가지 변수선택방법과의 보정 p-값 (adjusted p-value)이 0.01보다 작은 결과를 보여주었고 예측모형 요인의 경우 (스태킹-서포트벡터머신)과 (서포트벡터머신-그래디언트 부스팅)의 보정 p-값이 0.05보다 낮은 값을 보여주었다. (스태킹-랜덤포레스트) 비교의 경우 보정 p-값은 0.0518로 0.05보다 약간 큰 값을 보여주었다. 네가지 스태킹 모형간의 정확도의 통계적 유의미성을 확인하기 위하여 Cochran's Q검정을 실행한 결과 Q 통계 값은 Q = 14.8827이고 p 값은 1.92×10^{-3}이었다. Pairwise McNemar's 사후검정의 결과 (M1E-M2E), (M1E-M4E)의 보정 p-값이 0.05 보다

Table 2: The AUC values and the average AUC values

Name	Feature Selection	Prediction Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SE
M1E		Stacking	0.789	0.678	0.787	0.719	0.829	0.761	0.027
M1R	FS1: Hybrid	RF	0.776	0.694	0.776	0.714	0.797	0.751	0.020
M1S	(RF, ReliefF)	R–SVM	0.704	0.601	0.608	0.669	0.759	0.668	0.030
M1G		GBM	0.750	0.618	0.769	0.688	0.849	0.735	0.039
M2E		Stacking	0.704	0.656	0.784	0.689	0.655	0.697	0.024
M2R	FS2: Embedded	RF	0.638	0.639	0.769	0.676	0.617	0.668	0.027
M2S	(RF)	R–SVM	0.637	0.611	0.728	0.676	0.661	0.663	0.020
M2G		GBM	0.696	0.627	0.803	0.620	0.669	0.683	0.033
M3E		Stacking	0.700	0.666	0.642	0.687	0.745	0.686	0.016
M3R	FS3: Hybrid	RF	0.602	0.691	0.630	0.639	0.670	0.652	0.018
M3S	(LR, TuRF, Enet)	R–SVM	0.609	0.639	0.599	0.690	0.678	0.643	0.018
M3G		GBM	0.719	0.628	0.674	0.643	0.716	0.676	0.019
M4E		Stacking	0.703	0.685	0.780	0.711	0.754	0.727	0.018
M4R	FS4: Hybrid	RF	0.597	0.653	0.652	0.661	0.718	0.656	0.019
M4S	(LR, LASSO)	R–SVM	0.610	0.639	0.599	0.690	0.678	0.643	0.018
M4G		GBM	0.718	0.650	0.689	0.660	0.725	0.688	0.015

RF: random forest, GBM: gradient boost model, R-SVM: radial kernel support vector machine, knn: k-nearest neighbors, LR: logistic regression, Enet: elastic net

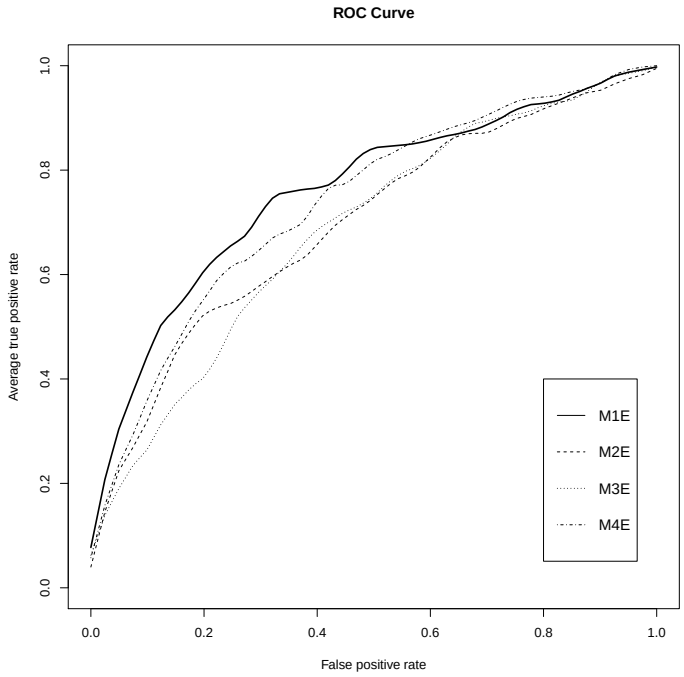
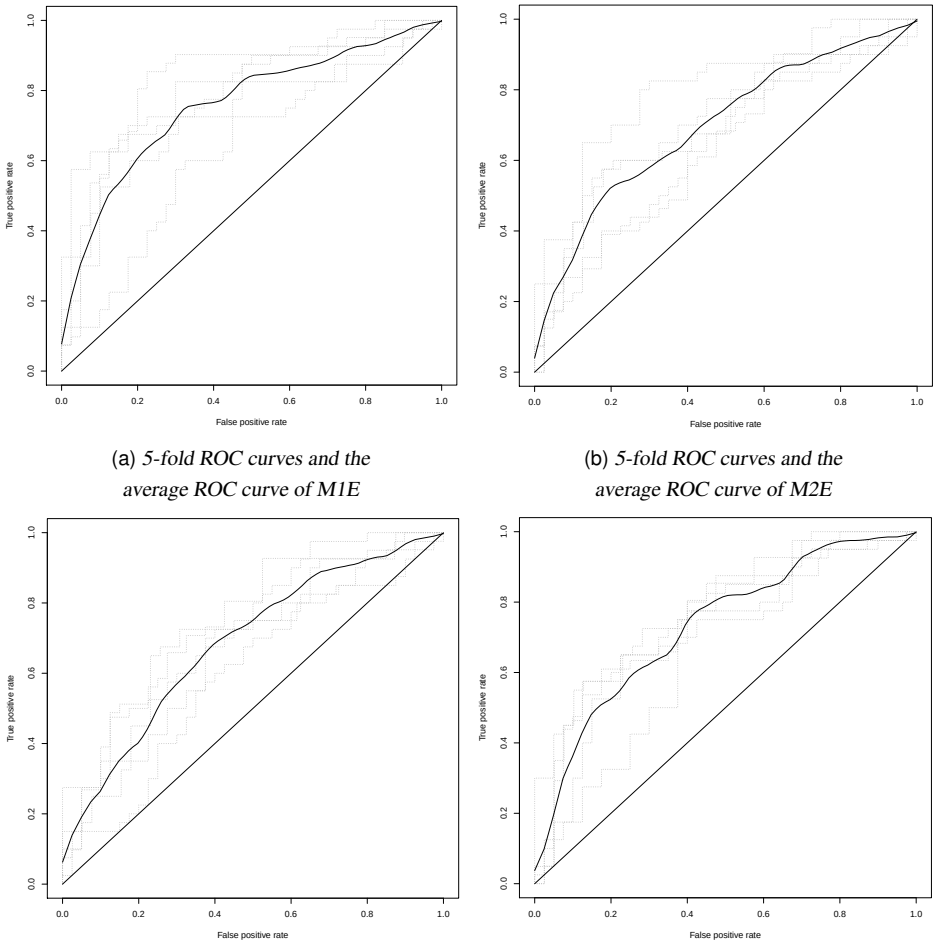


Figure 2: *The average ROC curves of the 4 stacking models.*



(a) *5-fold ROC curves and the average ROC curve of M1E*

(b) *5-fold ROC curves and the average ROC curve of M2E*

(c) *5-fold ROCs curve and the average ROC curve of M3E*

(d) *5-fold ROCs curve and the average ROC curve of M4E*

Figure 3: *5-fold ROC curves and the average ROC curves of 4 stacking models.*

작게 나왔으며 (M1E-M3E)의 보정 p-값은 0.062로 0.05보다는 약간 큰 값을 보여주었다. 분석 결과로부터 FS1 변수 선택 방법이 가장 좋은 성능을 보여주는 것을 알 수 있으며 예측 모형의 경우 스태킹 방법과 그래디언트 부스팅이 서포트벡터머신이나 랜덤포레스트보다 높은 성능을 보여주는 것으로 보인다. 성능이 높은 그래디언트 부스팅과 스태킹의 표준편차를 보면 스태킹의 표준편차가 더 낮은 경향이 있어 그래디언트 부스팅보다 스태킹 방법이 상대적으로 안정적인 모형으로 보인다.

5. 결론

맞춤형 의료 혹은 정밀 의료는 암과 같은 특정 질병 분야에서 활발히 연구되고 있으며 치료 방법 및 약물의 반응에 대한 예측모형 개발은 미래의 의학으로 불리고 있다. 이에 따라 환자의 유전자자료와 임상자료를 이용

하여 약물 반응을 예측하는 것은 점차 현대 의학의 필수적인 요소로 자리 잡아가고 있다. 기존의 유전통계학은 고전 통계를 적용하여 유전병 혹은 약물 반응의 유의미한 유전자 변수를 추출하는 것에 목적을 하고 있지만 약물유전체학의 한 분야는 인공지능, 데이터마이닝, 머신러닝과 같은 복잡한 블랙박스 모형의 활용으로 예측력이 높은 모형을 구축하는데 적용할 수 있다.

약물유전체 데이터는 많은 유전자 변수를 포함하는 고차원 데이터로 예측모형에 사용할 변수를 선택하는 과정이 필수이다. 각 질병이나 약물의 특성에 따라 효율적인 변수 선택 방법이 다를 수 있는 만큼 필터 방법, 래퍼 방법, 임베디드 방법이나 다양한 조합의 혼합 방법을 사용할 수 있다. 또한, 머신러닝의 예측모형 또한 더욱 다양한 방법을 사용해볼 수 있다.

이 논문에서는 400명의 뇌전증 환자의 약물 반응 예측을 위하여 로지스틱 회귀, ReliefF, TurF, 랜덤포레스트, LASSO의 혼합 방법과 같은 변수 선택 방법을 고차원의 약물유전체학 데이터에 적용하여 머신러닝 예측모형 구축에 효과적으로 이용할 수 있는 방안에 대해서 논의하였다. 또한 단일 모형을 적용했던 기존의 연구와는 다르게 여러 가지 머신러닝 모형을 교차검증 데이터에 적용하고 그 결과를 앙상블 하는 스태킹 모형을 적용해본 결과 통계적으로 유의미한 성능 향상을 볼 수 있었다. 이 과정에서 여러 가지 변수 선택 방법을 시도하여 모형 간의 성능을 비교하였으며, 같은 예측모형이라도 변수 선택 방법에 따라 성능에 유의미한 차이가 생길 수 있음을 확인하였다.

랜덤포레스트를 이용하는 변수 선택 방법은 두가지 단점을 가지는 것으로 알려져 있다. 첫 번째 단점은 취할 수 있는 값이 많은 변수를 선택하려는 경향이 있다는 것이고 두 번째 단점은 상관계수가 높은 변수를 선택하지 않으려는 경향이 있는 것이다. 이 논문에서 사용한 약물유전체 데이터의 경우 모든 유전자 변수는 0, 1, 2의 값을 취할 수 있어 첫 번째 단점을 극복하는 것으로 보인다. 특히 랜덤포레스트과 ReliefF를 이용한 변수 선택 방법이 높은 성능을 보여주었는데 이는 ReliefF가 랜덤포레스트의 두 번째 단점을 보완해주고 있는 것으로 보인다. 이 논문에서는 적은 표본 수로 인하여 변수 선택 과정의 초모수 선택 연구는 실행하지 않았으나 앞으로 더 많은 표본 수의 자료를 사용하여 다양한 변수 선택 방법 및 조합과 예측모형의 활용에 대한 연구가 필요할 것으로 보인다. 최근 Bommert 등 (2020)은 고차원 데이터를 사용하여 다양한 필터 방법의 비교 연구를 실행하였다. 유전체 자료의 경우 유전자 발현(gene expression) 자료를 사용한 변수 선택 방법 연구가 SNP 데어터보다 많이 실행되어 앞으로 더 많은 고차원 SNP 데이터에 대한 변수 선택 방법 및 비교 연구가 필요해 보인다.

References

Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, and Inouye M (2014). Accurate and robust genomic prediction of celiac disease using statistical learning, *PLoS Genetics*, **10**, e1004137.

Austin PC, Tu JV, Ho JE, Levy D, and Lee DS (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, *Journal of Clinical Epidemiology*, **66**, 398–407.

Boulesteix AL, Janitza S, Kruppa J, and Konig IR (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**, 493–507.

Bommert A, Sun X, Bischl B, Rahnenführer J, and Lang M (2020). Benchmark for filter methods for feature selection in high-dimensional classification data, *Computational Statistics and Data Analysis*, **143**, 106839.

Breiman L (1996). Stacked regressions, *Machine Learning*, **24**, 49–64.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Chen X and Ishwaran H (2012). Random forests for genomic data analysis, *Genomics*, **99**, 323–329.

Díaz-Uriarte R (2007). GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest, *BMC Bioinformatics*, **8**, 328.

Fisher RS, Acevedo C, Arzimanoglou A, *et al* (2014). Ilae official report: a practical clinical definition of epilepsy, *Epilepsia*, **55**, 475–482.

Genuer R, Poggi JM, and Tuleau-Malot C (2015). VSURF: an R package for variable selection using random forests, *The R Journal*, **7**, 19–33.

Ho DSW, Schierding W, Wake M, Saffery R, and O'Sullivan J (2019). Machine learning SNP based prediction for precision medicine, *Frontiers in genetics*, **10**, 267.

Jović A, Brkić K, and Bogunović N (2015). A review of feature selection methods with applications, *In 2015 38th international convention on information and communication technology*, electronics and microelectronics (MIPRO), 1200–1205.

Kang KW, Kim W, Cho YW, *et al* (2019). Genetic characteristics of non-familial epilepsy, *PeerJ*, **7**, e8278.

Kim MK, Moore JH, Kim, and Shin MH (2011). Evidence for epistatic interactions in antiepileptic drug resistance, *Journal of Human Genetics*, **56**, 71–76.

Kira K, and Rendell LA (1992). The feature selection problem: Traditional methods and a new algorithm. *In Aaai Press*.

Kononenko I (1994). Estimating attributes: analysis and extensions of RELIEF. *In European conference on machine learning*, 171-182, Springer, Berlin, Heidelberg.

Lee JY, Kim MK, and Kim W (2020). Robust linear trend test for low-coverage next-generation sequence data controlling for covariates, *Mathematics*, **8**, 217.

Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, and Liu H (2017). Feature selection: a data perspective, *ACM Computing Surveys (CSUR)*, **50**, 1–45.

López B, Torrent-Fontbona F, Vin as R, and Ferna ndez-Real JM (2017). Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction, *Artificial Intelligence in Medicine*, **85**, 43–49.

Mieth B, Kloft M, Rodríguez JA, *et al* (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies, *Scientific Reports*, **6**, 36671.

Montañez CAC, Fergus P, Montaez AC, Hussain A, Al-Jumeily D, and Chalmers C (2018). Deep learning classification of polygenic obesity using genome wide association study snps, *In 2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE.

Moore JH and White BC (2007). Tuning relieff for genome-wide genetic analysis. *In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, Springer.

Muštra M, Grgić M, and Delač K (2012). Breast density classification using multiple feature selection, *Automatika*, **53**, 1289–1305.

Nguyen TT, Huang JZ, Wu Q, Nguyen TT, and Li MJ (2015). Genome-wide association data classification and snps selection using two-stage quality-based random forests, *BMC Genomics*, Springer.

Romagnoni A, J egou S, Van Steen K, Wainrib G, and Hugot JP (2019). Comparative performances of machine learning methods for classifying crohn disease patients using genome-wide genotyping data, *Scientific Reports*, **9**, 1–18.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Van der Laan MJ, Polley EC, and Hubbard AE (2007). Super learner, *Journal of Clinical Epidemiology*, **6**.

Wei Z, Wang W, Bradfield J, *et al* (2013). Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease, *The American Journal of Human Genetics*, **92**, 1008–1012.

Wolpert DH (1992). Stacked generalization, *Neural Networks*, **5**, 241–259.

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

Received November 17, 2020; Revised December 28, 2020; Accepted February 1, 2021

약물유전체학에서 약물반응 예측모형과 변수선택 방법

김규환[a], 김원국[1,b]

[a]중앙대학교 통계학과, [b]중앙대학교 응용통계학과

요 약

약물유전체학 연구의 주요 목표는 고차원의 유전 변수를 기반으로 개인의 약물 반응성을 예측하는 것이다. 변수의 개수가 많기 때문에 변수의 개수를 줄이기 위해서는 변수 선택이 필요하며, 선택된 변수들은 머신러닝 알고리즘을 사용하여 예측 모델을 구축하는데 사용된다. 본 연구에서는 400명의 뇌전증 환자의 차세대 염기서열 분석 데이터에 로지스틱 회귀, ReliefF, TurF, 랜덤 포레스트, LASSO의 조합과 같은 여러 가지 혼합 변수 선택 방법을 적용하였다. 선택된 변수들에 랜덤포레스트, 그래디언트 부스팅, 서포트벡터머신을 포함한 머신러닝 방법들을 적용했고 스태킹을 통해 앙상블 모형을 구축하였다. 본 연구의 결과는 랜덤포레스트와 ReliefF의 혼합 변수 선택 방법을 이용한 스태킹 모형이 다른 모형보다 더 좋은 성능을 보인다는 것을 보여주었다. 5-폴드 교차 검증을 기반으로 하여 적합한 최적 모형의 평균 검증 정확도는 0.727이고 평균 검증 AUC 값은 0.761로 나타났다. 또한, 동일한 변수를 사용할 때 스태킹 모델이 단일 머신러닝 예측 모델보다 성능이 우수한 것으로 나타났다.

주요용어: AUC, 앙상블, 머신러닝, 랜덤포레스트, 스태킹

이 논문은 2019년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.
이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2018R1D1A1B07050012).

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: wkim@cau.ac.kr