

# 데이터마이닝

(Data Mining)

한국방송통신대학교  
정보통계학과 장영재 교수

6 강 /

---

# 양상블모형

# 목차

## 6. 양상불 모형

- 1) 양상불모형이란?
- 2) 양상불모형의 종류



# 1. 앙상블 모형이란?

# 앙상블모형이란?

## 1) 앙상블모형의 소개

- 앙상블 (ensemble) 모형이란 주어진 데이터를 이용하여 여러 개의 서로 다른 예측모형들을 생성한 후, 이러한 예측모형들의 예측결과들을 종합하여 하나의 최종 예측결과를 도출해 내는 방법
  - 앙상블모형은 목표변수의 형태에 따라 분류분석에도 사용가능하고, 회귀분석의 경우에도 사용가능
    - 현실적으로, 앙상블모형은 대부분 분류모형에서 사용되고 있는 실정
- 앙상블모형은 훈련데이터에서 B개 (B는 50, 100, 200 등의 값을 사용함)의 서로 다른 분류기의 예측을 종합하는 방법으로 주로 다수결 방법이 사용

# 앙상블모형이란?

## 1) 앙상블모형의 소개

- 다수결 방식에는 단순 다수결방식과 가중 다수결 방식

$$1\text{이라 예측} : \frac{1}{B} \sum_{i=1}^B \hat{T}_i > 0.5 \quad 1\text{이라 예측} : \sum_{i=1}^B w_i \hat{T}_i > 0.5$$

$$0\text{이라 예측} : \frac{1}{B} \sum_{i=1}^B \hat{T}_i \leq 0.5 \quad 0\text{이라 예측} : \sum_{i=1}^B w_i \hat{T}_i \leq 0.5$$

- 단, 각분류기의가중치인  $w_i$ 는 각분류기오류율의역수개념. 즉, 성능이우수한 분류기에가중치를더부여
- 대표적 분류앙상블 방법중에 배깅과 랜덤포레스트 방법은 단순 다수결방안을 사용하고, 부스팅 방법은 가중 다수결 방안을 사용

# 앙상블모형이란?

## 1) 앙상블모형의 성공요인

- 앙상블 방법이 성공하기 위해서는 앙상블 집합내에 구축되는 분류기 (classifier)  $T_i$  가 서로 유사하지 않고, 매우 다양
  - 앙상블모형에 사용되는 각각의 분류기  $T_i$ 는 같은 분류모형에 의해서 생성될 필요는 없음
  - 만약 나무모형 한 종류의 모형을 이용하여 B개의 분류기를 생성해 낸다면 다음과 같은 방법들을 활용해서 다양성을 확보

# 앙상블모형이란?

## 2) 앙상블모형의 성공요인

➤ 앙상블방법이 성공하기 위해서는 앙상블 집합내에 구축되는 분류기 (classifier)  $T_i$ 가 서로 유사하지 않고, 매우 다양

- 앙상블모형에 사용되는 각각의 분류기  $T_i$ 는 같은 분류모형에 의해서 생성될 필요는 없음
- 만약 나무모형 한 종류의 모형을 이용하여 B개의 분류기를 생성해 낸다면 다음과 같은 방법들을 활용해서 다양성을 확보



# 앙상블모형이란?

## 2) 앙상블모형의 성공요인

### (1) 데이터의 다양성

- 붓스트랩 방법은 반복이 있는 확률임의추출 방법을 의미하는데, 훈련데이터를  $L$  이라고 한다면 붓스트랩 방법은  $L_1, \dots, L_B$  개의 데이터를 생성
- 붓스트랩 방법은 배깅과 랜덤포레스트 앙상블 방법에서 사용

# 양상블모형이란?

## 2) 양상블모형의 성공요인

### (2) 나무모형 생성의 다양성

- 나무모형을 적합시킬 때, 분할방법에 변화를 주어 나무모형이 다양해지도록 만드는 방법
- 훈련데이터  $L$ 은 변동없이 그대로 있을지라도, 나무모형의 분할방법에 임의성을 가미
- 중간마디에서 분할후보점을 선발할 때에 전체 변수가 아닌, 임의의 변수의 부분집합중에서 선발하고, 그중에서 분할개선도를 최대화시키는 분할점을 탐색(랜덤포레스트 방법에서 이 방법을 활용)

## 2. 앙상블모형의 종류

## 2. 앙상블모형의 종류

### 1) 배깅 (Bagging) 방법

- 배깅은 bootstrap aggregating 의 약어로서, 훈련데이터로부터 붓스트랩 데이터를 B번 생성하여 각 붓스트랩 데이터마다 분류기를 생성한 후 그 예측 결과를 앙상블하는 방법
  - B는 앙상블의 크기라고 정의하며, 일반적으로 50을 많이 사용

## 2. 앙상블모형의 종류

### 1) 배깅 (Bagging) 방법

(1) 훈련데이터  $L = \{(x_i, y_i), i = 1, \dots, n\}$ 을 정의.

여기서  $x_i$ 는 입력변수 벡터이고  $y_i$ 는 목표변수

(2) L로부터 B개의 붓스트랩 데이터  $L_1, \dots, L_B$ 를 생성

(3) 각 붓스트랩 데이터  $L_1, \dots, L_B$ 에 대하여 분류기  $T_1, \dots, T_B$ 를 생성.  
흔히,  $T_1, \dots, T_B$ 는 분류나무

(4) B개의 분류기를 결합시켜 최종 예측모형  $\hat{f}(x)$ 를 생성

단,  $\hat{f}(x) = \arg \max_j \left\{ \sum_{b=1}^B I(T_b(x) = j), j = 1, \dots, J \right\}$

이고,  $x$ 는 예측하고자 하는 관찰치의 입력변수 벡터값.

분류모형인 경우  $\hat{f}(x)$ 는 다수결투표에 따라 집단을 분류하는 것과 같음



## 2. 앙상블모형의 종류

### 1) 배깅 (Bagging) 방법

- 배깅 방법은 불안정한 분류방법의 예측력을 획기적으로 향상시킨다고 알려져 있음 (데이터의 미세한 변동에도 적합된 모형의 결과가 많은 변화가 생기는 분류모형)
  - 분류나무 중에서도 가지치기를 사용하지 않은 최대나무를 사용하는 것이 더 예측정확도가 좋다고 알려져 있음
    - 가지치기를 사용하지 않은 경우에 더 불안정한 분류방법이 되기 때문

## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

- 부스팅 (boosting) 방법은 프로인드와 샤파이어 (Freund and Schapire, 1997)에 의해 개발된 분류앙상블 방법
  - 배깅과 마찬가지로 B개의 분류기를 생성하여 종합하는 방법인데, 분류기를 생성하는 방식과 종합하는 방식이 약간 상이
- 부스팅에 사용되는 분류기는 오분류율이 랜덤하게 예측하는 것보다 약간이라도 좋은 예측모형이기만 하면 효과가 있다고 알려져 있음
  - 예측력이 약한 분류모형들을 결합하여 강한 예측모형을 만드는 과정
    - 부스팅방법을 실행하는 알고리즘 중에서도 가장 많이 사용되는 아다부스트 (AdaBoost: adaptive boosting) 방법

## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

#### (1) 가중치 반영된 분류기 생성 방식

- 훈련데이터 관찰값들의 가중치를 반영한 분류기를 생성하는 방식
- CART 방법의 지니지수 공식은 각 관찰치들이 균등한 가중치를 갖는다는 것을 가정( $p(j|t) = \frac{n_j}{n}$ )
- 만약 관찰치에 따라서 더 높은 가중치를 갖는 경우가 있다면 지니지수의 공식에서  $p(j|t)$  값이 이를 반영
  - 각관찰치별가중치가  $w_i$  라고한다면,  $p(j|t) = \sum_i^n w_i I[y_i = j]$

## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

#### (1) 가중치 반영된 분류기 생성 방식

- 부스팅에서는 오분류가 발생한 데이터의 가중치를 높게 설정
- ∴ 분류기가 생성될때 가중치가 높은 관찰값에 대한 오분류를 줄이는 방향으로 적합될 것이기 때문에, 궁극적으로는 오분류율을 줄일 수 있기 때문 (현 분류기의 가중치는 직전 단계의 분류기에 의해 결정)

따라서 앙상블이 진행됨에 따라 지속적으로 오분류되는 관찰치는 더 높은 관심을 받게 되는 결과

- 가중치조정방식의아다부스트 (AdaBoost: adaptive boosting)

## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

(1) 훈련데이터  $L = \{(x_i, y_i), i = 1, \dots, n\}$  을 정의.

여기서  $x_i$  는 입력변수 벡터이고  $y_i$  는 목표변수

(2) 가중치  $w_i = \frac{1}{n}, i = 1, \dots, n$  을 초기화

(3)  $b=1, \dots, B$ 에 대하여 다음의 과정을 반복 수행

(a) 가중치  $w_i$ 를 이용하여 분류기  $T_b(x)$ 를 생성

(b)  $T_b(x)$ 를  $L$ 에 적용하여 각 데이터의 오분류 여부를 판별

(c)  $Err_b$ 를 다음과 같이 계산

$$Err_b = \sum_i^n w_i I[y_i \neq T_b(x_i)]$$



## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

(d) 분류기의 중요도  $\alpha_b$ 를 다음과 같이 계산

$$\alpha_b = \frac{1}{2} \log \frac{1 - Err_b}{Err_b}$$

(e) 가중치  $w_i$ 를 다음과 같이 업데이트

$$w_i = \frac{w_i}{Z} \exp\{\alpha_b I[y_i \neq T_b(x_i)] - \alpha_b I[y_i = T_b(x_i)]\}, i = 1, \dots, n$$

여기서  $Z$ 는  $w_i$ 의 합이 1이 되도록 만들기 위한 상수

## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

(4) 단계 3)에서 생성한 B개의 분류기를 결합하여 최종 예측모형

$\hat{f}(x)$ 를 생성

단,  $\hat{f}(x) = \arg \max_j \left\{ \sum_{b=1}^B I(T_b(x) = j), j = 1, \dots, J \right\}$

이고,  $x$  는 예측하고자 하는 관찰치의 입력변수 벡터값

- 분류모형인 경우  $\hat{f}(x)$  는  $\alpha_b$  라는 가중치를 반영한 가중투표 방법으로 집단을 분류하는 것과 같음

## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

(2) 표본추출에 의한 분류기 생성 방식

- 배깅에서는 균등확률 붓스트랩을 사용했다면, 부스팅에서는 가중확률 붓스트랩을 사용한다는 차이점
  - 표본추출 방식의 아다부스트 (AdaBoost : adaptive boosting) 방법

(1) 훈련데이터  $L = \{(x_i, y_i), i = 1, \dots, n\}$  을 정의.

여기서  $x_i$  는 입력변수 벡터이고  $y_i$  는 목표변수

(2) 가중치  $w_i = \frac{1}{n}, i = 1, \dots, n$  을 초기화

## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

(3)  $b=1, \dots, B$ 에 대하여 다음의 과정을 반복 수행

(a) 가중치  $w_i$ 를 이용하여 가중확률 붓스트랩으로  $L_b$ 를 생성

(b)  $L_b$ 를 이용하여  $T_b(x)$ 를 생성

(c)  $T_b(x)$ 를  $L$ 에 적용하여 각 데이터의 오분류 여부를 판별

(d)  $Err_b$ 를 다음과 같이 계산

$$Err_b = \sum_i^n w_i I[y_i \neq T_b(x_i)]$$

(e) 분류기의 중요도  $\alpha_b$ 를 다음과 같이 계산

$$\alpha_b = \frac{1}{2} \log \frac{1 - Err_b}{Err_b}$$

## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

(f) 가중치  $w_i$  를 다음과 같이 업데이트

$$w_i = \frac{w_i}{Z} \exp\{\alpha_b I[y_i \neq T_b(x_i)] - \alpha_b I[y_i = T_b(x_i)]\}, i = 1, \dots, n$$

여기서  $Z$ 는  $w_i$  의 합이 1이 되도록 만들기 위한 상수



## 2. 앙상블모형의 종류

### 2) 부스팅(Boosting) 방법

(4) 단계 3)에서 생성한 B개의 분류기를 결합하여 최종 예측모형  $\hat{f}(x)$ 를 생성

$$\text{단, } \hat{f}(x) = \arg \max_j \left\{ \sum_{b=1}^B I(T_b(x) = j), j = 1, \dots, J \right\}$$

이고,  $x$ 는 예측하고자 하는 관찰치의 입력변수 벡터값

- 분류모형인 경우  $\hat{f}(x)$ 는  $\alpha_b$ 라는 가중치를 반영한 가중투표 방법으로 집단을 분류하는 것과 같음
- 뿌리노드를 한번만 분할하는 나무인 스템프 (stump) 나무, 혹은 스템프 나무를 한번 더 분할하는 깊이가 2인 나무모형을 적합하여 사용하여도 훌륭한 예측 정확도를 보임

## 2. 앙상블모형의 종류

### 3) 랜덤포레스트(Random Forest) 방법

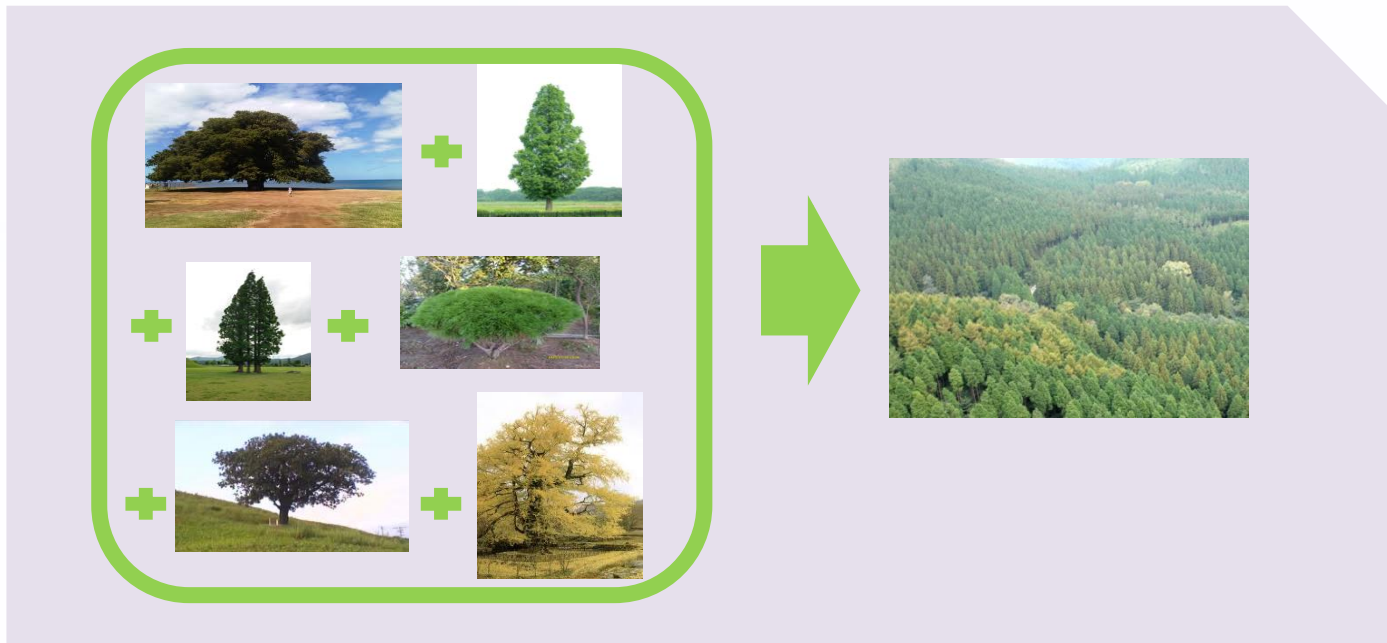
➤ 랜덤포레스트 방법은 배깅과 부스팅보다 더 정확한 예측력을 가지고 있다고 알려져 있음

- 특히 입력변수의 개수가 많을 때에는 그 효과가 극대화
- 랜덤포레스트에서 각 나무모형들을 생성할때는, 랜덤의 속성을 최대화 하기 위해 붓스트랩과 더불어 입력변수들에 대한 무작위 추출을 결합하는 방법을 취함

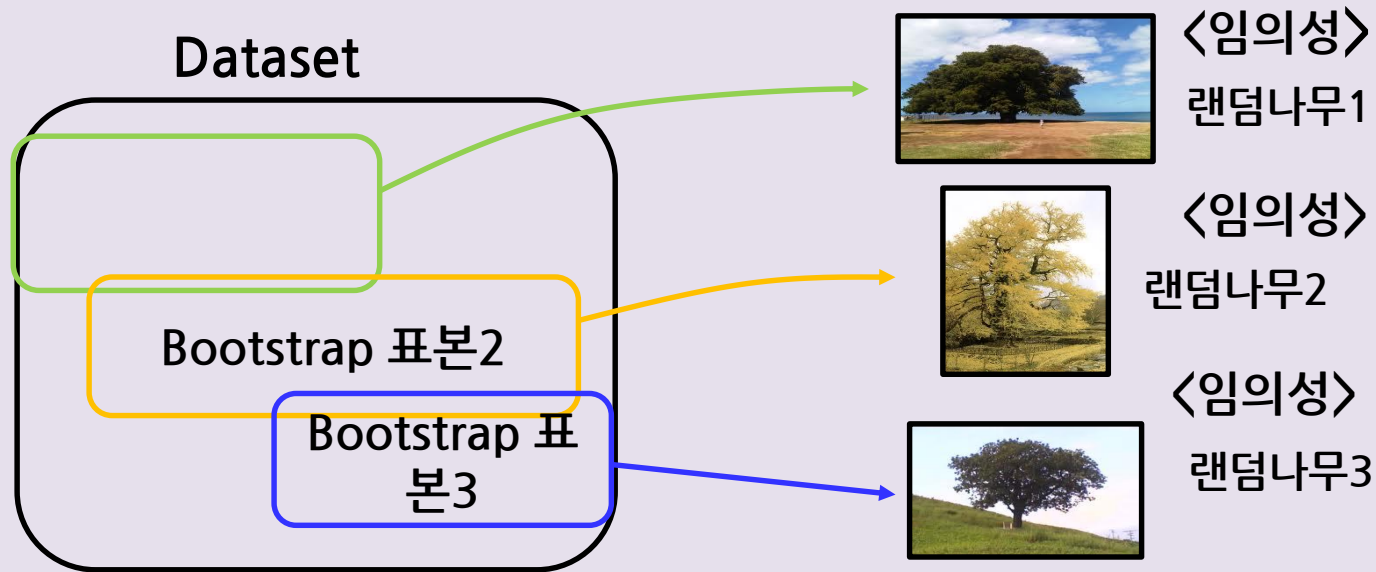
## 2. 앙상블모형의 종류

### ❖ 랜덤포레스트(Random Forest) 특징

- 임의성(Randomness) + 나무모형의 집합 = 예측력 향상



## 2. 앙상블모형의 종류



## 2. 앙상블모형의 종류

### ❖ 나무들의 결합

랜덤 나무1



랜덤 나무2



랜덤 나무3



선형  
결합

Random Forests





## 2. 앙상블모형의 종류

### 3) 랜덤포레스트(Random Forest) 방법

(1) 훈련데이터  $L = \{(x_i, y_i), i = 1, \dots, n\}$  을 정의.

여기서  $x_i^*$ 는 입력변수 벡터이고  $y_i$ 는 목표변수. 입력변수는  $p$ 개라 가정.  
즉,  $x_i = (x_{i1}, \dots, x_{ip})'$  이고 입력변수를 랜덤 추출한 벡터  $x_i^*$ 는  $x_i$ 에서  $M$ 개의 변수로 구성된 입력변수집합( $M \ll p$ )

(2)  $x_i$ 로부터  $B$ 개의 붓스트랩 데이터  $L_1, \dots, L_B$ 를 생성

(3)  $L_b$ 를 이용하여  $T_b(x)$ 를 생성. 단,  $T_b(x)$ 를 생성할 때 매 중간노드마다  $x_i$ 를 사용하지 않고,  $x_i^*$ 를 사용하여 나무모형을 생성

(4)  $b=1, \dots, B$ 에 대하여 (3)의 과정을 반복 수행

## 2. 앙상블모형의 종류

### 3) 랜덤포레스트(Random Forest) 방법

(5) 1) B개의 분류기를 결합시켜 최종 예측모형을 생성

$$\text{단, } \hat{f}(x) = \arg \max_j \left\{ \sum_{b=1}^B I(T_b(x) = j), j = 1, \dots, J \right\}$$

이고,  $x$ 는 예측하고자 하는 관찰치의 입력변수 벡터값

- 분류모형인 경우  $\hat{f}(x)$ 는 다수결투표에 따라 집단을 분류하는 것과 같음
- 랜덤포레스트 방법에서는 최적의  $M$ 값을 선택하는 방법은 알려져 있지 않지만, 브레이먼 (Breiman, 2001)은  $M = \sqrt{p}$ 로 하거나 혹은  $M = \frac{p}{3}$ 를 추천하며 또한  $M$ 의 최소값으로 5를 추천
- 배깅 방법과 마찬가지로 랜덤포레스트 방법은 가지치기를 사용하지 않은 최대나무를 사용하는 것이 더 예측정확도가 좋다고 알려져 있음

The background is a vibrant abstract composition featuring various shades of purple and blue. It includes large, soft-edged organic shapes, several circles with diagonal hatching patterns, and smaller circles with halftone dot patterns. A central white rounded rectangle contains the text.

**강의를 마쳤습니다.**  
다음시간에는...