

# 결정트리, 앙상블 학습, 랜덤 포레스트 실제 사용 사례 조사

---

9조 - 김한결, 위다빈

## 1. 결정트리

**결정트리(decision tree)란 ?**

-사용하는 문제 : 분류와 회귀 문제

**-학습방식 :**

데이터의 특성을 기반으로 분류나 예측을 위해 질문을 생성하며, 이를 통해 트리 구조를 형성하는 방식.

**-요약**

결정 트리를 학습 = 정답에 가장 빨리 도달하는 예/아니오 질문 목록을 학습

## 2. 실제 사례

대구시의 살피소

- 민원 : 민원인이 행정기관에 대하여 처분 등 특정한 행위를 요구
  - 민원행정 : 민원에 대응 행정에서 가장 기본적인지만 가장 해결하기 어려운 영역(문책, 민원인의 폭력등..)
  - 현재 상황 : 민원접수 건수가 매년 21% 이상 증가, 내용 또한 복잡 다양하게 진화
- 양질의 민원서비스 제공을 위한 방안 마련이 절실.
- 대구시의 살피소 : 공무원이 시민 불편사항을 먼저 찾아 처리하는 사전 예방 중심의 시정 건문정보시스템 (선제적 민원 대응)
- 문제 발생 : 처리건수가 매년 30% 이상 증가하는 등 양적 성장을 보이고 있지만 질적성장 (실질적으로 체감할 수 있는지)에 대한 확인필요
- 해결방안 모색 : 관리원과 대구시 민원행정 프로세스 혁신을 위한 빅데이터 분석 실시

→ 분석에 활용한 데이터 : 데이터로 2년간의 살피소 데이터, 시민이 직접 신청한 민원 현황, 유동인구 데이터 등 활용

→ 개발할 분석 모델 : 시민불편해소지수(시민불편 선제대응 지수) 개발, **취약지점 예측모델** (민원 빈발지점) 예측, 처리부서 자동지정

→ 취약지점 예측모델 :


유동인구 및 업종 분포 등 외부 데이터와 지도학습 기반 앙상블 학습 방법으로 다수의 의사 결정트리로부터 예측치를 모아 평균 또는 예측하는 랜덤 포레스트 머신러닝 알고리즘을 활용하여, 민원 취약지점을 96.2%의 높은 정확도로 예측하는 ‘취약지점 예측모델’을 개발,

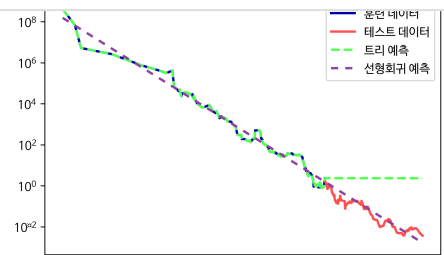
모델 적용 결과 동구 안심공업단지 주변 등 27곳이 향후 민원이 빈번히 발생할 지역으로 예측되어 해당 지역의 순찰 강화 등 효과적인 선제 대응이 가능해짐

## ▼ 참고한 곳

### 2.3.5 결정 트리


2.3.4 나이브 베이즈 분류기 | 목차 | 2.3.6 결정 트리의 앙상블 – 결정 트리decision tree는 분류와 회귀 문제에 널리 사용하는 모델입니다. 기본적으로 결정 트리는 결정에 다다르기 위해 예/아니

 <https://tensorflow.blog/파이썬-머신러닝/2-3-5-결정-트리/>



### 시민 불편, 인공지능으로 똑똑하게 살핀다

민원처리에 관한 법률(제2조 제1호)에 따르면, ‘민원’이란 ‘민원인이 행정기관에 대하여 처분 등 특정한 행위를 요구하는 것’이다. 이와 같이 간단히 정의되는 ‘민원’과 이에 대응하는 ‘민원행정’은

 <https://www.aitimes.kr/news/articleView.html?idxno=12573>



## 2. 앙상블 학습

### 1) 앙상블 학습(Ensemble Learning)이란 ?

-사용하는 문제 : 분류와 회귀 문제

-학습방식

### 1. 배깅 (Bagging):

- 동일한 알고리즘을 사용하지만 훈련 데이터의 서로 다른 서브셋에 대해 개별 모델을 훈련시킵니다.
- 중복을 허용한 무작위 샘플링 (부트스트래핑)을 통해 서브셋을 생성합니다.
- 모든 예측기의 예측을 평균내어 최종 예측을 생성합니다.
- 대표적인 예: 랜덤 포레스트

### 2. 부스팅 (Boosting):

- 순차적으로 여러 예측기를 훈련시키며, 이전 예측기의 오류를 다음 예측기가 수정하도록 합니다.
- 대표적인 예: AdaBoost, Gradient Boosting Machine (GBM), XGBoost.

### 3. 스택킹 (Stacking):

- 여러 가지 다른 모델들의 예측 결과를 새로운 '메타 모델'의 입력으로 사용하여 최종 예측을 생성합니다.

## -요약

분류 문제에 대해서는 다수결로, 회귀 문제에 대해서는 평균값으로 예측

## 실제사례

삼다수에서는 지하수 관리 시스템을 위해 앙상블 기법을 사용하고있음을 언급하였다.

제주개발공사에 따르면 딥러닝 인공지능뿐 아니라 최적 인공신경망(ANN)과 인공지능 앙상블 모델도 개발해 취수원 지하수위 예측 및 관리에 활용하고 있다고 한다.

구체적인 사용방법 및 특징을 공개하지 않아, 기사 내용을 바탕으로 조사한 내용은 아래와 같다.

1. **예측 모델링:** 지하수의 레벨, 흐름, 품질 등의 변화를 예측하기 위해 다양한 모델을 앙상블로 결합할 수 있습니다. 단일 모델보다 앙상블 모델은 더욱 정확한 예측을 제공할 수 있습니다.
2. **특성 중요도 분석:** 지하수의 특성 중 어떤 것이 가장 중요한지 판단하기 위해 랜덤 포레스트와 같은 앙상블 기법을 사용하여 특성 중요도를 평가할 수 있습니다.
3. **결측치 대체:** 결측치가 있는 데이터는 앙상블 기법을 사용하여 다른 관련 특성을 기반으로 예측하고 결측치를 대체할 수 있습니다.

4. **지하수 오염원 탐지**: 앙상블 기법은 다양한 데이터 소스(예: 위성 이미지, 수질 측정 데이터 등)를 결합하여 지하수 오염 원인을 식별하는 데 활용될 수 있습니다.
5. **시나리오 분석**: 다양한 환경 및 사용 패턴에 따른 지하수 상태의 변화를 예측하기 위해 앙상블 기법을 활용할 수 있습니다.
6. **알림 및 경보 시스템**: 앙상블 모델을 사용하여 지하수 수준이나 품질에 관한 임계치를 초과할 가능성이 있는 시점을 예측하고, 관리자나 사용자에게 알림을 보낼 수 있습니다.

**지하수 관리 시스템을 위한 특징**은 지하수 수준, 수질, 기후, 토양 유형, 토지 이용, 추출량, 지하 구조, 지역 특성, 인구 밀도, 역사적 데이터 등이 될 수 있다.

▼ 참고한 곳

<https://www.fnnews.com/news/202310111758565273>

### 3. 랜덤 포레스트

#### 1. 랜덤 포레스트 (Random forest)란 ?

-사용하는 문제 : 분류와 회귀 문제

- 학습방식 :

##### 1. 부트스트랩 샘플링 (Bootstrap Sampling)

- 원래 데이터셋에서 무작위로 중복 허용하여 샘플을 추출합니다. 이렇게 추출된 샘플로 각 결정 트리가 학습됩니다.
- 이 방법으로 인해 각 트리는 조금씩 다른 데이터를 바탕으로 학습되며, 이는 랜덤 포레스트의 다양성을 증가시킵니다.

##### 2. 특성의 무작위 선택

- 각 결정 노드에서 모든 특성을 고려하는 것이 아니라, 무작위로 선택된 일부 특성만을 고려하여 최적의 분할을 찾습니다.
- 이 방법 또한 모델의 다양성을 증가시키며, 과적합을 방지하는 효과도 있습니다.

##### 3. 결정 트리 학습

- 부트스트랩 샘플과 선택된 특성을 바탕으로 결정 트리를 학습합니다.

- 일반적으로 트리의 깊이에 제한을 두지 않거나, 가지치기를 수행하지 않습니다. 이는 각 트리가 과적합될 가능성이 있지만, 전체 랜덤 포레스트의 앙상블 방식으로 인해 과적합 문제가 완화됩니다.

#### 4. 예측

- 회귀의 경우, 각 트리의 예측값의 평균을 결과로 합니다.
- 분류의 경우, 각 트리의 예측 클래스를 투표 방식으로 집계하고 가장 많은 투표를 받은 클래스를 최종 예측 결과로 합니다.

#### 5. 특성 중요도 평가

- 랜덤 포레스트는 각 특성의 중요도를 계산할 수 있습니다. 일반적으로 특성 중요도는 각 트리에서 특성을 사용하여 데이터를 분할할 때의 평균 감소량(불순도 감소 또는 평균제곱오차 감소)을 기반으로 합니다.

#### -요약 :

랜덤 포레스트는 여러 결정 트리를 조합해 데이터의 부분 집합과 특성을 무작위로 학습하여 고정밀도와 일반화 성능을 제공하는 앙상블 기법입니다.

## 2. 실제 사례

정부에서 교차로 접근부 추돌사고 중 전치 3주 이상의 중상에 해당하는 심각사고의 요인을 찾기 위해 다양한 머신러닝 기법을 적용하였음.

이 중 랜덤 포레스트 또한 존재. 정부에서 구체적으로 어떻게 사용하였는지에 대한 내용은 존재하지 않지만, 기사에 나온 ‘인적 측면, 차량 측면, 도로환경적 측면에서 총 27개의 다양한 변수를 수집해 분석’이라는 문장을 보아, 총 27개의 특징을 통해 아래 과정에서 적절하게 활용하였을 것으로 예상됨.

아래는 사용사례를 찾으며 분석한 랜덤 포레스트의 실제 활용 방법임.

### 1. 데이터 수집 및 전처리

교통사고 데이터를 수(사고의 세부 정보, 운전자 정보, 환경 요인, 차량 정보 등)

타겟 변수를 정의(전치 3주 이상의 중상 사고는 1로, 그렇지 않은 사고는 0으로 레이블링)

### 2. 특성 선택

사고 발생 시의 환경 (날씨, 도로 상태, 시간대 등)

운전자 정보 (나이, 성별, 경험 등)

차량 정보 (차량 유형, 연령, 속도 등)

교통량, 신호등 상태, 보행자 유무 등 교차로에 관한 정보

### 3. 랜덤 포레스트 모델 훈련

데이터를 훈련 세트와 테스트 세트로 분할

랜덤 포레스트 모델을 훈련 데이터로 학습

### 4. 특성 중요도 평가

랜덤 포레스트는 각 특성의 중요도를 평가하는 기능을 제공함. 따라서 이를 통해 심각사고와 가장 관련이 깊은 요인을 파악할 수 있음

### 5. 모델 평가

테스트 세트를 사용하여 모델의 성능을 평가합니다. 정확도, 정밀도, 재현율, F1 점수, ROC 곡선 등 다양한 메트릭을 사용하여 모델을 평가할 수 있음.

#### ▼ 참고한 곳

<https://www.moneys.co.kr/news/mwView.php?no=2023101616464474576>