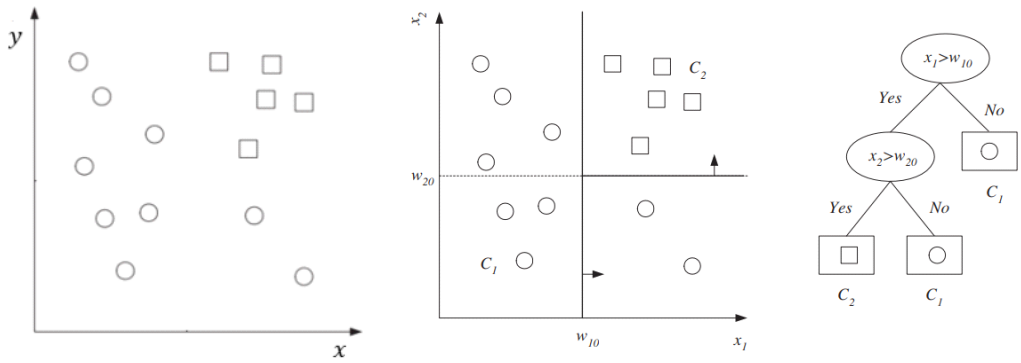
의사결정나무분석(Decision Tree)과 앙상블 방법(Ensanble)

IT기술의 발달로 데이터의 크기, 다양성, 생성속도가 증가함에 따라 과거 할 수 없었던 대량의 데이터를 학습하여 미래를 예측하는 머신러닝이 많은 주목을 받고 있다. 머신러닝은 학습 데이터에 레이블이 있는 경우와 그렇지 않은 경우에 따라 각각 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 구분한다. 레이블(Label)이라는 것은 학습 데이터의 속성을 연구자가 분석하고자 하는 관점에 따라 정의하는 것이다. 예를 들어, 우리 주변에 있는 사물을 찍은 사진 중에서 어떤 사물이 있는지 구분하는 과제가 있다면, 갖고 있는 사진을 학습 데이터라 하고 사진 속에 있는 사물의 이름을 레이블이라고 한다. 레이블은 연구자가 사진을 보고 정의한 것이기 때문에, 컴퓨터 입장에서 레이블된 사진을 읽어 학습하는 것이 사람에게 지도(Supervised)를 받아 학습하는 것이 된다. 반면 학습 데이터에 레이블이 없다면 컴퓨터가 사람으로부터 지도받지 않았기 때문에 비지도학습이라 한다. 지도학습에는 분류와 회귀가, 비지도학습에는 군집이 대표적이다. 그중 지도학습의 의사결정나무(Decision Tree)는 대표적인 분류 모형으로서 해석력이 높으며, 나무 형태의 시각화된 그래프가 학습 결과로 도출되기 때문에 사회 전 분야에 널리 활용되고 있다.

의사결정나무는 기본적으로 데이터가 위치한 좌표평면의 반복적 직교분할 방법을 사용한다. 분할을 위한 최적 분기의 결정은 분할된 영역 안에 있는 데이터의 '불순도(Impurity)' 지표를 기준으로 그 값이 가장 낮은 지점을 분기로 선택한다. 이러한 불순도 지표 중 지니 계수(Gini index)가 대표적으로 활용된다. 지니 계수는 불순도를 측정하는 지표로서, 값이 낮을수록 각 집합의 순수도가 높은 것으로 볼 수 있다. D영역 에 m개의 레이블을 갖는 데이터가 각각 속할 확률을 p_i라 하면, 지니 계수는 다음과 같이 계산한다.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

의사결정나무가 2차원 상의 데이터 ○(C_1), □(C_2) 를 분류하기 위한 x, y좌표평면을 분할하는 과정을 살펴보면 <그림>과 같다. 먼저 xy좌표평면에 해당하는 모든 x, y값으로 직교분할했을 경우의 지니 계수를 산출한다. x값은 W_{10}, y값은 W_{20}이 각각 불순도 지수가 가장 낮은 값으로 나타났을 경우, W_{10}과 W_{20}의 불순도 지수를 비교하게 된다. 불순도를 의미하는 지니 계수는 W20보다 W_{10}으로 분할 하였을 때 더 낮기 때문에, 불순도가 더 낮은 W_{10}지점을 최초의 직교분할 분기로 선택하게 된다. 다음으로 x값이 W_{10}보다 작은 영역에 속해있는 집합은 이미 충분히 순수하기 때문에 더 이상 분할 하지 않고, x좌표값이 W_{10}보다 큰 집합만 분할하게 된다. 이때에도 역시 이 영역에 속해있는 모든 x, y값의 불순도 지수가 계산되어 가장 불순도 지수가 낮은 W_{20} 값이 분기로 선택된다.



$$Gini(w_{10}) = [1-(\frac{7}{7})^2-(\frac{0}{7})^2] \times \frac{7}{14} + [1-(\frac{2}{7})^2-(\frac{5}{7})^2] \times \frac{7}{14} = 0.204$$

$$Gini(w_{20}) = [1-(\frac{5}{5})^2-(\frac{0}{5})^2] \times \frac{5}{14} + [1-(\frac{4}{9})^2-(\frac{5}{9})^2] \times \frac{9}{14} = 0.317$$

이렇게 분할한 좌표평면은 거꾸로된 나무 형태의 그래프로 시각화되어 나타난다. 좌표평면에서의 분할 분기인 W_{10}, W_{20}는 노드라 한다. 각 위치에 따라 명칭이 달라지는데, 나무의 최상단 노드는 루트(Root), 중간 분기는 내부 노드(Internal Node), 가장 말단에 위치한 더 이상 분할되지 않는 노드는 잎 노드(Leaf Node)라 한다. 잎 노드의 속성값이 바로 그룹을 대표하는 레이블(예: ○, □ 또는 C_1, C_2)이다. 각 노드를 속성이라고도 하며 각 속성들은 분기를 결정하는 속성값을 가진다. 상황은 이러한 일련의 속성값에 따라 결정된다. 예를 들면, $x>W_{10}$ → $x>W_{20}$ → □(C_2)가 하나의 상황이며, 규칙이다. 이러한 상황은 환자의 진료기록을 토대로 발병여부를 유추하는 경우, 대출자의 채무 불이행 여부, 그리고 대출을 받고자 하는 사람의 신용등급을 평가하는 경우에 많이 활용되고 있다. 이와 같은 의사결정나무분석은 모든 데이터가 완벽하게 분류될 때까지 더 작은 부분으로 영역을 분할하면서 트리가 무한정 커질 수 있다. 이렇게 완전히 성정한 나무는 해당 데이터에 종속되어 일반화시키기 어려운 과적합(Over Fitting) 모형이 될 수 있다. 이러한 과적합을 방지하기 위해 새로운 데이터에도 적용할 수 있도록 나무의 크기를 적절하게 줄여주는 가지치기(Pruning)를 통해 결과를 일반화하게 된다.

의사결정나무는 얼마나 나무를 성장시킬지, 그리고 성장한 나무를 얼마나 어떤 식으로 가지치기할지 등을 연구자가 판단해야 하는 어려움이 있다. 또한 다른 머신러닝 방법에 비해 설명력이 높은 반면 예측력이 다소 떨어지며, 데이터의 하위 집합이 조금만 바뀌어도 분기 값이 달라지는 모형 안정성이 떨어진다는 단점이 많이 꼽힌다.

이와 같은 의사결정나무모형의 단점을 해결하기 위해 앙상블 기법이 대안으로 활용되고 있다. 앙상블은 여러 개의 의사결정나무를 만든 뒤, 그 결과를 종합하는 기법이다. 특히 지도학습 중 분류의 경우 여러 개의 트리 결과를 표결로 종합해 다수결로 선택된 레이블을 따르기 때문에, 나무 한 개가 잘못된 분류를 할 수는 있어도, 과반수 이상의 나무가 잘못된 분류를 하지 않는 이상 정확도는 낮아지지 않는다. 이러한 기법 중 대표적인 것이 배깅과 부스팅이다. 배깅(Baagging)은 학습 데이터에서 나무를 여러개 생성하고 난 뒤, 결과를 종합하는 방법이며, 부스팅(Boosting)은 첫 번째 나무를 생성하면서 잘못 분석한 데이터에 가중값을 주어 두 번째 나무가 해당 데이터를 분석할 때 더 주의 깊게 분석할 수 있도록 하는 방법이다. 배깅은 각각의 나무가 병렬적으로 연관되어 있는 반면, 부스팅은 각각의 트리가 서로 연쇄적으로 연결되어 있는 특징이 있다.

* 참 고 문 헌 *

1. 학술 논문 :

- 이유나(2019). Cost-Sensitive Learning을 활용한 심뇌혈관 질환 발생 예측 모형 개발. 충북대학교 석사학위 논문. 6쪽.

- 유진은(2015). 랜덤 포레스트: 의사결정나무의 대안으로서의 데이터 마이닝 기법. 교육평가연구, 28(2), 427-448.

- 그림 출처: Alpaydin, E. (범어디자인연구소 번역) (2018). 머신러닝 쉽게 이해하기. 유엑스리뷰.

2. 책 : 김의중(2016). 인공지능, 머신러닝, 딥러닝, 입문. 위키북스.

3. 사이트 : https://wikidocs.net/39491