과제 #2

결정 트리, 앙상블 학습, 랜덤 포레스트의 실제 사용 사례 조사

<8조> - 202084009 김승현 / 202004197 김은미

의사 결정 트리

- 의사결정트리는 일련의 분류 규칙을 통해 데이터를 분류, 회귀하는 지도 학습 모델 중 하나.
- 결과 모델이 Tree 구조를 가지고 있기 때문에 Decision Tree 라는 이름을 가진다.

1. 의학 분야

   [사용 사례 1] 한림대동탄성심병원, 인공지능으로 요관결석 치료법 결정

   요관결석은 요로계에 결석이 생겨 소변의 흐름이 원활하지 않게 되고 그 결과 칼로 찌르는 듯한 극심한 통증을 유발하는 질환이다.

   요관결석 치료는 크게 3가지로 나뉘는 데 이 중 충격파를 통해 몸 밖에서 결석을 분쇄하는 **체외충격파쇄석술은 외부 상처 없이 결석을 제거할 수 있어 가장 선호**된다. **그러나 체외충격파쇄석술만으로 모든 결석을 치료할 수 없고 치료 전 정확한 결과를 예측하기 어려워 환자가 시간과 비용만 낭비하고 결국 수술과 같은 다른 치료를 추가로 받는 경우**가 생긴다.

   한림대동탄성심병원과 한림대학교 연구팀은 **인공지능을 활용해 요관결석 환자에게서 체외충격파쇄석술 성공여부를 사전에 확인할 수 있는 예측모델을 개발**했다.

   연구팀은 2012년 10월부터 2016년 8월까지 한림대학교동탄성심병원에서 요관결석으로 체외충격파쇄석술을 받은 환자 791명을 분석했다. 전체 환자 중 509명(64.3%)은 체외충격파쇄석술로 결석 제거에 성공했으며 282명(35.7%)은 실패했다. 이후 두 환자군의 상세한 데이터를 인공지능을 통해 분석했다.

   연구에서 활용한 인공지능 모델은 의사결정트리 알고리즘이다. 프로그램이 데이터를 통해 학습하는 기계학습을 통해 새로운 데이터를 입력하게 되면 최적의 판단이나 결과를 예측하는 모델을 만드는 것이다.

   인공지능 분석 결과 환자의 나이, 성별, 결석의 상태 등 총 15가지 요인을 분석하는 체외충격파쇄석술 성공률 예측 모델이 만들어졌다. 예측을 위한 15가지 요인은 요

관 결석 환자가 병원에서 처음 검사하게 되는 소변검사, 혈액검사, CT 3가지 검사만으로 확인 가능하다.

연구팀은 이 모델을 다시 요관결석 환자 100명에게 적용했고 그 결과 92.29%의 정확도로 체외충격파쇄석술 성공여부를 예측하는 것으로 나타났다.

체외충격파쇄석술 성공률 예측모델이 개발됨에 따라 요관결석 환자들은 시간과 비용을 절약할 수 있게 됐다. 초기검사를 통해 체외충격파쇄석술로 결석 제거가 가능한지 확인이 가능하기 때문에 체외충격파쇄석술 실패에 따른 비용을 낭비하지 않을 수 있고 다음 치료까지 환자의 고통이 길어지는 것도 막을 수 있다.

출처 : 메디게이트 뉴스(https://www.medigatenews.com)

[사용 사례 2] 머신러닝 기반 암 사망예측 모델 개발

서울의대와 국립암센터 연구팀은 **폐암 치료 후 암 생존자들의 생활 습관 및 삶의 질 정보를 활용해 머신러닝 기반의 사망예측 모델을 개발했으며, 이를 통해 5년 후 암 생존자의 사망을 보다 정확하게 예측하는데 성공**했다.

연구팀은 잘 알려진 폐암 예후 인자(연령, 성별, 병기요인, 종양의 특성 등)외에도 삶의 질과 생활습관 정보(불안, 우울, 삶의 질, 긍정적 성장 및 과체중)들이 실제로 암 생존자들의 5년 이후의 생존예측력을 높일 수 있는지를 중점적으로 연구했으며, 이에 대한 예측정확도를 높이고자 **머신러닝 알고리즘을 적용**했다.

폐암 생존자들의 사망률을 평가하기 위해 컴퓨터가 예제를 통해 학습하는 데 도움이 되는 지도학습 알고리즘 중, **다섯 가지 유형의 머신러닝 알고리즘을 테스트**했다. 그 이후 각각의 모델에 대한 예측 성능을 비교했다.

다섯 가지 유형의 알고리즘은 하나의 모델을 학습시켜 사용하는 **의사결정나무(decision tree)과 로지스틱회귀분석(logistic regression)**, 가능한 임의의 결과를 반영하는 여러 개의 나무 모양 모델을 결합한 **랜덤포레스트(random forest), 배깅(Bagging), 아다부스트(Adaptive Boosting)** 등 이다.

폐암 치료 후 암 생존자들의 생활 습관 및 삶의 질 정보를 활용해 개발 된 사망 예측 모형은 기존의 잘 알려져 있는 예후 요인인 연령, 성별, 종양의 특성 등만 활용한 모델의 사망 예측보다 훨씬 더 정확했다고 연구팀은 밝혔다. 또한 **다양한 머신러닝기법을 적용함으로써 암 사망에 대한 예측력을 보다 높일 수 있다는 것을 확인**했다.

출처 : 의학신문(http://www.bosa.co.kr)

---

2. 데이터 분류

[사용 사례] 의사결정트리 기반 기계학습을 이용한 식품교환표 식품군 분류 모델 연구

기존 식품과 웹 크롤링으로 찾은 식품 데이터에 대해 기계학습으로 **식품군을 분류하여 식품교환표를 갱신하기 위한 의사결정트리 기반의 기계학습 모델**을 제안한다.

기존의 식품군을 바탕으로 새롭게 추가되는 식품을 분류하기 때문에 식품의 트렌드를 반영한 식품교환표 구성이 가능하다. 제안 모델로 식품을 분류한 결과, 식품교환표의 식품군에 대한 정확도가 97.45%로 나타났으며, 본 식품 분류 모델은 병원, 요양원 등에서 식단 구성에 활용도가 높을 것으로 전망된다.

출처 : 의사결정트리 기반 기계학습을 이용한 식품교환표 식품군 분류 모델 (한국컴퓨터정보학회, 2022 / 김지윤, 김종완)

---

3. 가격 분석

[사용 사례] 의사결정트리(Decision Tree)를 활용한 글로벌 부동산 가격 분석 연구

의사결정트리 모형을 이용하여 글로벌 부동산 가격하락과 관련성이 높은 변수를 분석하였다.

분석기간은 1976년부터 2018년까지이며, 58개국 불균형 국가패널자료(연도별)를 사용하여 3단계 분석을 수행하였다.

분석결과 글로벌 부동산 가격하락 이벤트와 연관성이 높은 중요도 순서 상위 15개 설명변수를 도출하였다. 주요 15개 변수에 대한 글로벌 부동산 가격하락 이벤트 기간 평균과 2016년 주요국의 해당 지표들을 비교한 결과, 한국은 2016년 전후 부동산 가격의 급격한 조정 가능성이 다른 국가와 비교해 상대적으로 낮았던 것을 확인할 수 있었다.

연구를 통해 단기 부동산 가격 전망에 유용한 참고자료가 될 수 있을 것으로 기대한다.

출처 : 의사결정트리(Decision Tree)를 활용한 글로벌 부동산 가격 분석 (사회과학연구 제29권 제1호 / 2022김경훈)

- **이 외에도 금융 서비스, 제조업, 마케팅, 환경 모니터링, 리스크 분석, 에너지 관리 등에 활용되고 있다.**

앙상블 학습

- 여러 개의 분류기(Classifier)를 생성하고 그 예측을 결합함으로써 보다 정확한 예측을 도출하는 기법

    1. 의료 분야

        [사용 사례] 오지서도 서울 유명 병원과 같은 진료 가능"…ETRI, 닥터AI 개발

        한국전자통신연구원(ETRI, 원장 김명준)은 여러 병원에 구축된 의료지능을 통합해 환자의 현재 상태를 정밀하게 분석하고 미래건강을 합리적으로 예측하는 인공지능 주치의 '닥터 AI(Dr. AI)'를 개발했다.

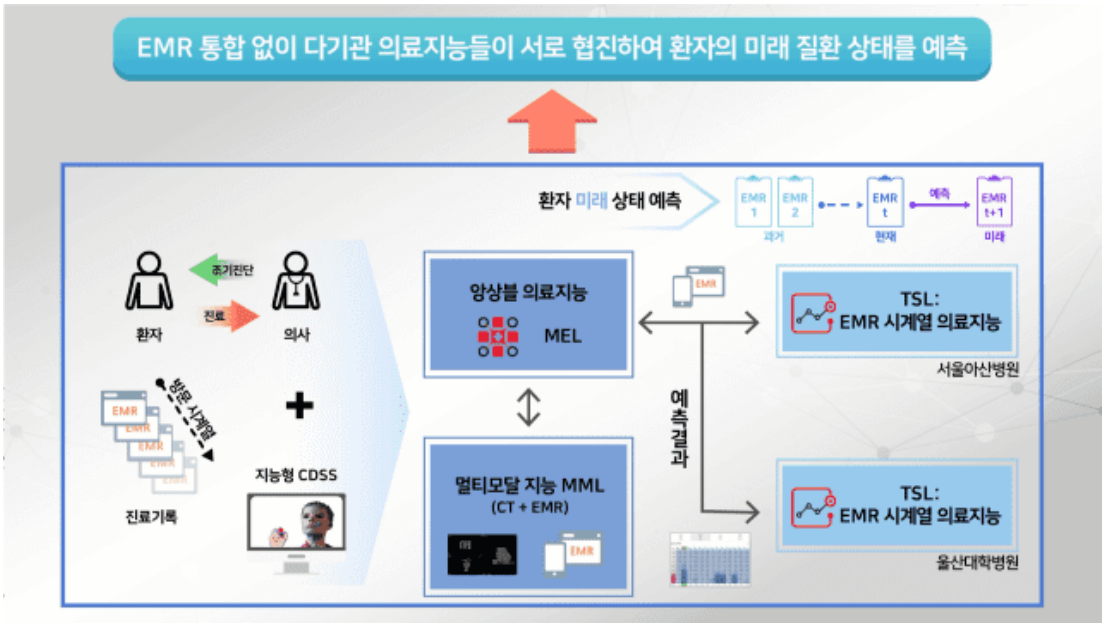
        '닥터 AI'는 각 병원의 환자 진단기록인 전자의무기록(EMR)을 통합하는 대신 각 병원의 EMR 기반 의료지능을 동시에 활용하는 방식(앙상블)으로 의사의 진료를 돕는다. 즉, 민감정보에 직접 접근하지 않으면서 다른 기관의 의료 데이터를 공동 활용하는 효과가 있다.

        코로나19로 비대면 진찰 필요성과 함께 의료 데이터를 학습해 환자를 분석하고 진단하는 의료 인공지능 기술이 새롭게 떠오르고 있는데, 의료지능 시스템을 구축하기 위해서는 각 병원의 EMR을 직접 통합해 환자별 의료 데이터를 수집 및 축적해야 하지만 현재 우리나라 의료법 및 제도상 한계가 있다.

        이에 ETRI가 개발한 '닥터 AI'는 EMR을 통합하는 대신 각 병원의 EMR 기반 의료지능을 동시에 활용하는 방식(앙상블)으로 진료를 돕는다.

        이론적으로는 이 서비스가 활성화되면 백령도 같은 오지에서도 서울의 유명 병원과 같은 의료 서비스를 받을 수 있다.

        ETRI가 개발한 '닥터 AI'의 핵심기술은 ▲앙상블 의료지능(기관별 예측 추세 및 오차 분석) ▲시계열 EMR 의료지능(예측 근거 및 건강상태 분석) ▲멀티모달 의료지능(의료 데이터 학습) 등이다.



출처 : https://zdnet.co.kr/view/?no=20211027105111

---

2. 금융 서비스

   [사용 사례] 교보 생명, 보험 사기 잡는 '인공지능 시스템' 특허

   최근 **보험사기** 적발 금액이 연간 1조원에 육박하는 등 관련 피해 사례가 속출하고 있다.

   이러한 가운데 교보생명(사장 편정범)이 **인공지능을 활용해 보험 사기 패턴을 수시로 재학습해 변종 보험사기를 검출하는 시스템**을 마련해 관심을 모으고 있다.

   이 시스템은 고객 이름과 병명·병원 등 다양한 의료 정보를 수집하는 '입원 데이터 수집부'와 수집 데이터를 기반으로 보험 사기 검출 작업을 수행하는 ▲입원 데이터 전처리부 ▲군집화 처리부 ▲학습 데이터 생성부 ▲보험 사기자 판정부로 구성된다.

   보험사기 검출 시스템은 입원 데이터 수집부에서 해당 데이터뿐만 아니라 보험 사기 내역, 위험 요소 등을 종합 관리한다.

   시스템이 본격 가동되면 입원 데이터 전처리부에서 데이터를 '쓸만한 자료'로 바꾸는 작업이 수행된다. 예를 들어 3일 이내 동일 병원을 재입원한 경우 입원기간을 합쳐 수정하는 작업을 거친다. 일반 보험 청구자를 사기 가해자로 오인하는 경우의 수를 줄이기 위함이다.

이후 본격적으로 보험사기 집단과 일반인 집단을 비교분석하는 작업(군집화)이 이어진다. 이 과정에서 '클러스터링 알고리즘'이 사용된다. 데이터들 사이에서 일정 구간 밀도가 가장 높은 데이터, 즉 출연 빈도가 높은 데이터를 집단으로 묶는 작업이다. 암·심장질환 같은 중대 질병을 '레드 존'으로 묶는 것을 예시로 들 수 있다.

인공지능 활용 방식은 머신러닝 기법 중 '앙상블 모델 기법'이 쓰인다. 앙상블 모델은 집단 분류 모델 여러 개를 생성해 최적의 답을 찾아내는 기법이다. **보험 사기 데이터를 하나의 데이터 뭉치로 볼 것이 아니라 여러 개의 비교군으로 설정해 이전보다 높은 신뢰성을 담보할 수 있다는 장점**이 있다.

교보생명 관계자는 "시스템 도입으로 보험회사 손실도 최소화하고 보험료 인상 등 소비자 피해도 줄일 수 있을 것"이라고 말했다.

출처 : https://www.bizwnews.com/news/articleView.html?idxno=59832

---

3. **[사용 사례] 영작문 자동채점 시스템 개발에서 학습데이터 부족 문제 해결을 위한 앙상블 기법 적용의 효과**

   적고 편향되기 쉬운 학습데이터와 이를 이용한 여러 평가영역에 대한 학습모델을 생성해야하기 때문에, 이를 위한 적절한 기계학습 알고리즘을 결정하기 어렵다.

   이러한 문제를 앙상블학습을 통해 완화할 수 있음을 실험한다.

   실제 중, 고등학교 학생들을 대상으로 시행된 단문형 영작문 채점 결과를 학습데이터 개수와 편향성을 조절하여 실험하였다. 학습데이터의 개수 변화와 편향성 변화의 실험 결과, 에이다부스트 알고리즘을 적용한 결과를 투표로 결합한 앙상블 기법이 다른 알고리즘들 보다 전반적으로 더 나은 성능을 나타냄을 실험을 통해 나타내었다.

   출처 : 영작문 자동채점 시스템 개발에서 학습데이터 부족 문제 해결을 위한 앙상블 기법 적용의 효과 (정보과학회논문지 제42권 제9호, 2015 / 이경호, 이공주)

4.

   [사용 사례] 모바일 멀티모달 센서 정보의 앙상블 학습을 이용한 장소 인식

   시각, 음향, 위치 정보를 포함하는 멀티모달 센서 입력 정보로부터 사용자가 위치한 장소의 환경 정보를 학습하고 기계학습 추론을 통해 장소를 인식하는 방법을 제안한다.

이 방법은 음영 지역에서의 정확도 감소나 추가 하드웨어 필요 등 기존 위치 정보 인식 방법이 가지는 제약을 극복 가능하고, 지도상의 단순 좌표 인식이 아닌 논리적 위치 정보 인식을 수행 가능하다는 점에서 해당 위치와 관련된 특정 정보를 활용하여 다양한 생활편의를 제공하는 위치 기반 서비스를 수행하는데 보다 효과적인 방법이 될 수 있다.

제안하는 방법에서는 스마트폰에 내장된 카메라, 마이크로폰, GPS 센서 모듈로부터 획득한 시각, 음향, 위치 정보로부터 특징 벡터들을 추출하여 학습한다. 이때 서로 다른 특성을 가진 특징 벡터들을 학습하기 위해 각각의 특징 벡터들을 서로 다른 분류기를 통해 학습한 후, 그 결과를 기반으로 최종적인 하나의 분류 결과를 얻어내는 앙상블 기법을 사용한다.

실험 결과에서는 각각의 데이터를 따로 학습하여 분류한 결과와 비교하여 높은 성능을 보였다. 또한 사용자 상황인지 기반 서비스의 성능 향상을 위한 방법으로서 제안하는 모델의 스마트폰 앱 구현을 통한 활용 가능성에 대해 논의한다.

출처 : 모바일 멀티모달 센서 정보의 앙상블 학습을 이용한 장소 인식 (정보과학회 컴퓨팅의 실제 논문지 제21권 제1호, 2015 / 이충연, 이범진, 온경운, 하정우, 김홍일, 장병탁)

- **이 외에도 자동차 산업, 환경 모니터링, 텍스트 및 음성 처리, 보안 분야 등에 활용되고 있다.**

랜덤 포레스트

- 랜덤 포레스트는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 분류 또는 평균 예측치(회귀)를 출력함으로써 동작한다.

1. 의료 분야

   [사용 사례] 뇌영상 AI, 자폐스펙트럼장애 진단한다

   자폐스펙트럼장애는 아동의 약 1~2%에서 발병하는 신경발달장애다. 주로 사회적 관계형성의 어려움, 정서적 상호작용의 문제, 반복적 행동과 제한된 관심 등이 특징이다.

그동안 자폐스펙트럼장애 진단은 발달과정에서의 이상 행동이나 표현을 관찰한 후 증상평가를 통해 내려졌다. **이 진단법은 전문가 간에 일치도는 높으나, 관찰자의 주관이 개입할 여지가 있고 발병원인과 연관성을 파악할 수 없다는 한계**가 있었다.

서울대병원 김붕년 교수(장수민 전임의)·한양대병원 이종민 교수(김인향 교수) 공동연구팀은 **2015년 5월부터 2019년 9월까지 58명의 자폐스펙트럼장애 환자군과 48명의 대조군을 대상으로 MRI 뇌영상 기반 머신러닝 AI알고리즘을 통해 진단 구분 능력을 평가한 연구 결과**를 17일 밝혔다.

참여자의 연령대는 3~6세였으며, 자폐군에는 저기능 환자(IQ 70미만)만 포함됐다.

머신러닝 알고리즘은 랜덤포레스트 등 기계학습을 적용해 분류기 형태로 구축됐다. 분류의 매개변수는 △T1강조 MRI 영상(대뇌 회백질의 특성을 정량적으로 측정) △확산텐서영상(대뇌 백질의 특성을 정량적으로 측정) △다중 MRI(T1강조 MRI·확산텐서영상을 조합해 측정)가 사용됐다.

연구팀은 매개변수별로 T1강조 MRI 모델, 확산텐서영상 모델, 다중 MRI 모델을 나눴다. 이후 머신러닝 AI알고리즘을 통해 자폐군과 대조군으로 진단 구분하는 능력을 각각 평가했다.

그 결과 다중 MRI 모델에서 정확도 88.8%, 민감도 93.0%, 특이도 83.8%로 높은 진단 구분 능력을 보여줬다. 특히 다중 MRI 모델의 정확도는 T1강조 MRI(78.0%)와 확산텐서영상(78.7%)을 단독으로 활용했을 때보다 10%p 향상된 것으로 나타냈다.

또한 자폐스펙트럼장애를 진단하는 가장 중요한 영상지표는 후두엽의 피질두께, 소뇌각의 확산도, 후측 대상회 연결도로 밝혀졌다.

김붕년 교수(소아청소년정신과)는 “이번 연구로 발달지연이 심한 영유아 자폐스펙트럼장애 환자를 생물학적 지표에 근거해 진단함에 있어, **기계학습을 통한 다중 MRI의 활용이 유용하다는 것을 확인할 수 있었다**”고 말했다.

출처 : https://www.medifonews.com/news/article.html?no=164982

---

2. [사용 사례] **의사결정나무 및 랜덤포레스트 분류 모델을 이용한 교량 안전등급 예측**

   의사결정나무 및 랜덤포레스트 분류 모델을 사용하여 교량의 안전등급을 예측하는 방법을 제안하였다. 일반국도상 교량 8,850개를 대상으로 해당 모델들을 혼동행렬, 균형 정확도, 재현율, ROC 곡선 및 AUC와 같이 여러가지 평가 지표를 통해 분석한 결과 전반적으로 랜덤포레스트가 의사결정나무보다 더 나은 예측 성능을 보유하였다. 특히 랜덤포레스트 중 랜덤언더 샘플링 기법은 노후도가 비교적 커서 유지관리에 주의를 기울여야 하는 C, D등급 교량에 대해 재현율 83.4%로 다른 샘플링 기법

들보다 예측 성능이 더 뛰어난 것으로 나타났다. 제안된 모델은 최근 점검이 실시되지 않은 교량들의 신속한 안전등급 파악 및 효율적이고 경제적인 유지관리 계획 수립에 유용하게 활용될 수 있을 것으로 기대된다.

출처 : 의사결정나무 및 랜덤포레스트 분류 모델을 이용한 교량 안전등급 예측 (대한토목학회논문집 제43권 제3호(통권 제228호), 2023 / 홍지수, 전세진)

---

3. [사용 사례] **RIDS: 랜덤 포레스트 기반 차량 내 네트워크 침입 탐지 시스템**

   CAN(Controller Area Network) 버스에서 해킹에 의한 공격을 탐지하기 위한 랜덤 포레스트 기반 침입 감지 시스템(RIDS: Random Forest-Based Intrusion Detection)을 제안한다. RIDS는 CAN 버스에서 나타날 수 있는 전형적인 세 가지 공격, 즉 DoS(Denial of Service) 공격, Fuzzing 공격, Spoofing 공격을 탐지하며, 데이터 프레임 사이의 시간 간격과 그 편차, 페이로드끼리의 해밍 거리와 그 편차의 네 가지 파라미터를 사용하여 공격을 판단한다. RIDS는 메모리 중심 방식의 아키텍쳐를 가지며 노드의 정보를 메모리에 저장하여 사용하며 트리의 개수와 깊이만 조절하면 DoS 공격, Fuzzing 공격, Spoofing 공격을 모두 탐지할 수 있도록 확장이 용이한 구조로 설계되었다. 시뮬레이션 결과 RIDS는 정확도 0.9835, F1 점수 0.9545로 세 가지 공격을 효과적으로 탐지할 수 있었다.

   출처 : RIDS: 랜덤 포레스트 기반 차량 내 네트워크 침입 탐지 시스템 (전기전자학회논문지 제26권 제4호, 2022 / 이대기, 한창선, 이성수)

- 이 외에도 랜덤 포레스트는 금융, 판매 및 마케팅, 환경 및 농업, 제조업 등에 활용되고 있다.