연구보고서 2017-46

기계학습(Machine Learning)기반 사회보장 빅데이터 분석 및 예측모형 연구

오미애 · 최현수 · 김수현 · 장준혁 · 진재현 · 천미경

KIHASA
한국보건사회연구원
Korea Institute for Health and Social Affairs

【책임연구자】
오미애 한국보건사회연구원 연구위원

【주요저서】
2016년 보건복지통계정보 통합 관리 및 운영
한국보건사회연구원, 2016(공저)

보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안
한국보건사회연구원, 2014(공저)

【공동연구진】

최현수 한국보건사회연구원 연구위원
김수현 한국과학기술연구원 책임연구원
장준혁 한양대학교 융합전자공학부 교수
진재현 한국보건사회연구원 전문연구원
천미경 한국보건사회연구원 연구원

연구보고서 2017-46

기계학습(Machine Learning) 기반
사회보장 빅데이터 분석 및 예측모형 연구

발 행 일 2017년 12월
저 자 오 미 애
발 행 인 김 상 호
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1층~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1994년 7월 1일(제8-142호)
인 쇄 처 ㈜현대아트컴
가 격 6,000원

ISBN 978-89-6827-464-0 93510

발간사

현 정부에서 여러 분야의 정보를 연계하여 인공지능(AI)을 통해 새로운 부가가치를 창출하고자 하는 노력은 4차 산업혁명 시대가 도래한 것과 무관하지 않다. 4차 산업혁명의 핵심 기술은 인공지능과 빅데이터로, 대규모로 축적된 데이터에 대한 기계학습(Machine Learning) 기반 자가 진화를 통한 알고리즘 성능의 지속적인 강화가 모든 분야에서 중요한 원천 기술이라고 할 수 있다.

기계학습(Machine Learning)은 AI 의 한 분야로 데이터를 바탕으로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야로 예측(Prediction)에 탁월한 성과를 나타낸다.

이 연구에서는 현재 국내에서 생산되고 있는 사회보장 빅데이터의 특성을 파악하고 기계학습(Machine Learning) 통계기법을 연구함으로써, 사회보장 빅데이터 분석에 적합한 기계학습 기반 예측모형을 설계하고 근거기반(evidence-based) 연구에 적용할 수 있는 방법론을 제시하였다. 기계학습 기반 예측 모형은 데이터의 활용 가능성을 높이고 다양한 분석을 가능하게 하여 새로운 가치를 도출하는 데 기여함으로써 학술적·정책적 함의를 제시할 수 있는 기반을 제공할 수 있다는 데 그 의의가 있다.

본 연구는 정보통계연구실 빅데이터연구팀 오미애 연구위원의 책임하에 최현수 연구위원, 진재현 전문연구원, 천미경 연구원, 한국과학기술연구원 김수현 책임연구원, 한양대학교 융합전자공학부 장준혁 교수에 의하여 수행되었다. 연구진은 바쁘신 중에도 본 보고서를 읽고 조언을 주신

본원의 고제이 부연구위원, 건국대학교 권성훈 교수에게 감사의 뜻을 전한다.

본 보고서의 결과는 우리 연구원의 공식적인 견해가 아니라 연구진의 의견임을 밝혀둔다.

2017년 12월

한국보건사회연구원 원장

김 상 호

목 차

표 목차

그림 목차

Abstract

A Study on Social security Big Data Analysis and Prediction Model based on Machine Learning

Project Head · Oh, Miae

In the age of the Fourth Industrial Revolution, we predict the future by linking various kinds of data. It creates new knowledge and values and forms a framework for establishing and enforcing better policies.

The core technology of the Fourth Industrial Revolution is artificial intelligence and big data, and the continuous enhancement of algorithm performance through Machine Learning based on large scale accumulated data is an important source technology in all fields. Machine Learning is a field of AI that develops algorithms and techniques that enable computers to learn based on data. It is a core technology in various fields such as image processing, image recognition, speech recognition, and Internet search. The results show excellent performance in predication.

In this study, we study the characteristics of Social Security Big Data which is being produced in Korea and study Machine Learning statistical techniques. We will design a Machine Learning-based prediction model suitable for Social Security

Co-Researchers: Choi, Hyunsoo · Kim, Suhyun · Chang, Joonhyuk · Jin, Jaehyun · Cheon, Mikyung

Big Data analysis and present a methodology that can be applied to evidence-based research. Machine Learning-based prediction model can provide a basis for presenting academic and policy implications by contributing to the utilization of data and making various analysis possible to derive new value. Machine learning can be applied to the field of social policy in which vast amounts of data are accumulated, and Machine Learning-based statistical analysis will be able to approach the predictable customized welfare.

요 약

1. 연구의 배경 및 목적

4차 산업혁명 시대에는 다양한 종류의 데이터를 연계하여 미래를 예측함으로써 새로운 지식과 가치를 창출하며, 더 나은 정책을 수립, 집행하는 과정을 마련하는 것이 핵심이다.

4차 산업혁명의 핵심 기술은 인공지능(AI)과 빅데이터로, 대규모로 축적된 데이터에 대한 기계학습(Machine Learning) 기반 자가 진화를 통한 알고리즘 성능의 지속적인 강화가 모든 분야에서 중요한 원천 기술이라고 할 수 있다. 기계학습(Machine Learning)은 AI 의 한 분야로 데이터를 바탕으로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이며, 이미지 처리, 영상인식, 음성인식, 인터넷 검색 등의 다양한 분야의 핵심 기술로 예측(Prediction)에 탁월한 성과를 나타낸다.

이 연구에서는 현재 국내에서 생산되고 있는 사회보장 빅데이터의 특성을 파악하고 기계학습(Machine Learning) 통계기법을 연구함으로써, 사회보장 빅데이터 분석에 적합한 기계학습 기반 예측모형을 설계하고 근거기반(evidence-based) 연구에 적용할 수 있는 방법론을 제시하고자 한다. 기계학습 기반 예측 모형은 데이터의 활용 가능성을 높이고 다양한 분석을 가능하게 하여 새로운 가치를 도출하는 데 기여함으로써 학술적·정책적 함의를 제시할 수 있는 기반을 제공할 수 있다. 기계학습은 방대한 데이터가 쌓이고 있는 사회정책 분야에 활용 가능하며 기계학습 기반 통계분석으로 예측 가능한 맞춤형 복지에 가까이 다가설 수 있게 할 것이다.

2. 주요 연구결과

본 연구에서는 사회보장 빅데이터와 기계학습 개념을 정의하고, 기계학습 기법이 사용된 다양한 활용 사례를 살펴보고, 최신 기계학습 기법이라고 할 수 있는 딥러닝(Deep Learning)에 대해 알아보았다. 기계학습 기법에 대한 이론적 배경 소개와 각 방법론별 장·단점을 비교하였다. 기계학습 기법을 적용한 예측모형을 평가할 수 있는 평가 방법을 살펴보고 한국복지패널을 이용하여 예측모형 시뮬레이션을 통해 여러 방법론을 비교·적용해보았다. 마지막으로, 기계학습 기법 적용과 관련된 이슈로 알고리즘의 책무성, 데이터 주도 정책 중심으로 다루었다.

본 연구의 구성에 따라 주요내용을 요약하면 다음과 같다.

2장에서는 사회보장 빅데이터 및 기계학습 개념을 정의하고 활용 사례 및 딥러닝 기법의 동향을 분석하였다. 사회보장 빅데이터의 개념은 사회보장의 정의에 따라 사회보장이 포괄하는 영역의 범위와 관련되며, 이를 토대로 “사회보장 분야에서 생산 및 구축된 행정데이터 또는 빅데이터”로 정의될 수 있다. 기계학습은 컴퓨터가 사전에 프로그램되어 있지 않고 데이터로부터 패턴을 학습하여 새로운 데이터에 대해 적절한 작업을 수행하는 일련의 알고리즘이나 처리 과정을 의미한다. 기계학습이 가장 큰 강점을 보이는 분야는 패턴인식에서 비롯되는 이미지 및 영상, 음성 인식과 자연어처리이다. 구글, 아마존, AiCure, 3D Robotics 등 많은 기업들이 기계학습을 적극적으로 이용하여 기업 가치를 극대화하고 있다. 딥러닝은 기계학습의 한 분야로, 기존의 머신러닝과 차이점이 있다면, 기계학습에서 기계가 학습하기 위해 주어진 데이터에서 특징을 추출하는 과정에 여전히 사람이 개입하지만 딥러닝은 주어진 데이터를 그대로 입력데이터로 활용한다는 점이다. 사람이 생각한 특징을 학습하는 것이 아니라 데이

터 자체에서 중요한 특징을 기계 스스로 학습하기 때문에 딥러닝을 end-to-end machine learning이라고 한다. 딥러닝의 가장 큰 장점 중 하나로는 기존의 shallow learning으로는 풀 수 없는 복잡한 문제들을 풀 수 있다는 것이다. 딥러닝의 DNN, CNN, RNN, LSTM, RL, VAE 등의 기법을 살펴봄으로써 딥러닝의 전반적인 이해를 돕고자 하였다.

3장에서는 기계학습 통계기법 비교 연구를 위해 반응변수(종속변수)가 범주형인 경우 적용할 수 있는 Logistic regression, Shrinkage method인 Elastic Net, Decision Tree, Random Forest, Boosting, Support vector machines 방법에 대해 살펴보았다.

4장에서는 기계학습 기반 예측모형 평가방법 연구를 위해, 모형 평가 방법의 개념을 정의하고 모형 선택기준을 살펴보았다. 모형 평가는 하나의 자료를 분석할 때, 여러 가지 통계모형을 상정하여 분석하는 것이 바람직하며, 여러 가지 모형 중 자료를 최적으로 설명하는 모형을 선택하는 것이 좋다. 모형의 평가란, 예측(prediction)을 위해 만든 모형이 임의의 모형(random model)보다 예측력이 우수한지, 고려된 다른 모형들 중 어느 모형이 가장 우수한 예측력을 보유하고 있는지를 비교, 분석하는 과정이라고 할 수 있다. 모형을 평가하기 위해 사용되는 선택기준으로는 오분류표(오분류율)를 이용한 정확도, 민감도, 특이도와 ROC, AUC, Lift 방법을 중심으로 살펴보았다.

5장에서는 다양한 기계학습 방법론을 활용한 모형별 비교 평가를 위한 모의분석을 위해 복지수급 예측모형 분석 DB를 구축하였다. 2015년 7월 기초생활보장제도 급여체계 개편에 따라 맞춤형 급여체계로 변화된 상황을 고려하여 기초생계급여의 수급 여부에 영향을 미치는 다양한 가구 특성 및 소득·재산 정보들을 중심으로 한국복지패널 11차 가구 패널로 DB를 구축하였다. 특히 2015년 말부터 보건복지부, 사회보장정보원, 한국

보건사회연구원이 공동으로 연구 및 시스템 구축을 통해 전국 지방자치단체의 일선 읍면동을 통해 시행 중인 복지사각지대 발굴관리시스템의 주요 사각지대 관련 요인 정보와 유사한 박탈 관련 문항에 대한 조사 결과들을 반영하여 분석 DB를 구축하였다. 반응변수를 독립표본으로 간주하여 t-test를 실시한 결과, 가구형태, 주거위기, 경제상태, 생활여건 대부분의 설명변수가 유의수준 0.05하에서 유의하였다. 모형 평가를 위해 10 fold CV 방법을 이용하여 4장에서 기술한 방법론을 중심으로 분류기준에 따라 정확도, 특이도, 민감도를 살펴본 결과 부스팅이 전반적으로 우수한 성능을 보였다. ROC, AUC, Lift 의 %Response 통계량 역시 모두 앙상블 기법인 부스팅과 랜덤 포레스트의 결과가 제일 좋음을 알 수 있었다.

6장에서는 기계학습과 관련된 이슈들을 정리하고 기계학습 방법이 적용되기 위해 필요한 부분이 무엇인지를 살펴보았다. 그리고 데이터 주도 혁신 및 정책 추진을 통한 새로운 가치 창출의 필요성이 대두되고 있는 현 시점에 고려해야 하는 사항들을 제시함으로써 기계학습이 가져다 줄 수 있는 장점을 극대화하고자 하였다.

3. 결론 및 시사점

이 연구에서는 앞서 정의한 사회보장 빅데이터와 유사한 분석DB를 구축하여 딥러닝을 포함한 다양한 기계학습 통계기법을 적용하고 여러 모형 선택기준으로 최적의 예측모형을 살펴보았다는 점에서 기존 연구와 차별성이 있다. 분석DB에서는 기계학습 기법인 부스팅의 결과가 제일 좋은 성능을 보여주었다. 딥러닝은 이미지 및 언어 분석에서는 탁월한 성능을 보여주지만, 분석DB에서는 부스팅과 랜덤 포레스트의 결과보다 좋지

않음을 알 수 있다. 이는 모든 데이터에서 하나의 기계학습 기법이 좋은 성능을 보이는 것은 아니라는 것을 의미한다. 부스팅 기법은 대부분의 경우 좋은 예측력을 보여주지만, 분석 데이터의 속성에 따라서 다른 기계학습 방법이 더 좋을 수 있다. 사회보장 빅데이터가 이미지 및 영상 데이터인 경우에는 부스팅 방법보다는 딥러닝 방법이 더 적합할 수 있다. 따라서, 기계학습 기법을 적용하기 위해서는 주어진 문제를 해결하기 위해 구축된 사회보장 빅데이터의 속성을 우선 이해해야 할 것이다. 복지사각지대 발굴 시스템의 경우 복지수급 대상자와 복지수급 비대상자의 인구학적 특성, 경제적 특성, 주거 특성 등에 대한 정보가 필요하다. 위기아동 발굴시스템은 학대 아동과 일반 아동의 부모 특성, 아동 특성이 중요한 정보일 것이다. 보건의료 분야의 영상데이터의 경우에는 이상자료(abnormal data)의 특성을 파악할 수 있어야 한다.

보건복지 분야는 다른 분야와 마찬가지로 데이터 주도 혁신 및 정책 추진을 통한 새로운 가치 창출의 필요성이 대두되고 있다. 과거 데이터 분석은 현상에 대한 단순한 정보를 제공하거나 원인을 진단하는 것이었다면, 4차 산업혁명 시대에는 미래 예측을 통해 최적화(optimization)된 해법을 도출, 제공하는 것이 중요하다. 보건복지 분야에서 구축되는 다양한 데이터를 기반으로 보건의료 빅데이터와 사회보장 빅데이터를 구축하여 공개하고, 기계학습 방법론을 활용한 예측을 통해 시간, 공간, 인간에 대해 최적화된 문제 해결 방안으로 데이터 주도 정책을 추진함으로써 우리 삶의 다양한 분야에서 전반적으로 활용하고 새로운 경제적·사회적 가치를 창출할 수 있다.

데이터 기반의 과학적 정책 결정이 이루어지는 예측형, 예방형 데이터 주도 보건복지정책을 추진하기 위해서는 정부를 포함한 보건복지 분야 전반에서 데이터 중심의 의사 결정 문화로 변화할 필요가 있다. 특히 보

건복지 관련 빅데이터 생산·관리 인프라를 기반으로 다양한 정책 변화에 따른 국민의 삶의 질 변화와 관련된 다양한 데이터를 구축하고 이러한 데이터를 활용한 예측 결과에 기반을 둔 정책 목표 설정과 정책 설계, 환류 데이터 분석을 통한 정책 수정과 실천이 병행되어야 한다. 또한, 보건복지 분야의 데이터 활용도 제고를 위해서는 데이터 거버넌스와 데이터 공유 플랫폼에 대한 정부 지원과 사회적 신뢰 기반 구축이 필수적이다.

이러한 데이터 주도 정책이 뒷받침된다면, 기계학습 분야는 더 발전하고 우리 삶의 보건복지 분야에도 매우 중요한 영향을 미칠 것이다.

*주요용어: 기계학습, 사회보장 빅데이터, 예측모형

제 1 장 서론

1 서론

제1절 연구배경 및 목적

1. 연구배경 및 필요성

2017년 10월 구글 딥마인드는 '네이처'에 인간 지식 없이 바둑 정복하기(Mastering the game of Go without human knowledge) 제목의 논문으로 알파고제로(AlphaGo Zero)를 공개했다. 이는 데이터에 대한 학습 없이 스스로 진화하는 강화학습을 보여준 사례이다. 우리나라에서는 2016년 3월 알파고로 인해 인공지능(Artificial Intelligence, AI)에 대한 일반 국민의 이해와 관심이 높아졌다. 2017년 초 다보스 포럼과 정부의 지능정보사회 종합대책이 발표되고 5월 대통령 선거를 거치면서 4차 산업혁명에 대한 정책적 관심 역시 집중되고 있는 상황이다.[1)]

4차 산업혁명은 빅데이터(Big data)와 인공지능을 핵심으로 하는 지능정보기술이 우리 삶의 다양한 분야에 보편적으로 활용됨으로써 새로운 가치가 창출되고 발전하는 사회를 의미한다. 4차 산업혁명(4IR: the Fourth industrial revolution)이란, 정보통신기술(ICT) 융합으로 인한 혁명의 시대를 의미하며, 2016년 1월 다보스에서 개최된 세계경제포럼(World Economic Forum)에서 제시된 이후 전 세계적으로 크게 주목받고 있다. 1784년 영국에서 시작된 증기기관과 기계화에 의한 1차 산업혁명, 1870년 전기 이용과 노동력 분화를 통한 대량생산이 본격적으로

1) 연구배경 및 필요성은 최현수, 오미애(2017a)의 일부 내용을 수정·보완하여 재구성함.

시작된 2차 산업혁명에 이어 3차 산업혁명은 1969년 정보기술(IT)과 인터넷이 이끈 정보화 및 자동생산시스템이 주도한 것이라면, 최근 모든 영역에서 가장 큰 이슈가 되고 있는 4차 산업혁명은 인공지능과 빅데이터를 중심으로 급격하게 진행되는 우리 미래의 혁신적인 변화를 의미한다(최현수, 오미애, 2017b, p. 1). 이러한 4차 산업혁명의 핵심은 인공지능, 로봇공학, 사물인터넷(IoT), 무인 운송 수단(자율주행차량, 무인항공기), 3차원 인쇄(3D 프린팅), 나노기술 등 6대 분야에서 나타나고 있는 기술 혁신이다. 4차 산업혁명은 물리적, 생물학적, 디지털 세계를 빅데이터에 입각해 통합하고 사회, 경제, 산업 등 모든 분야에 영향을 미치는 다양한 신기술로 설명될 수 있다.

이처럼 4차 산업혁명 시대에는 다양한 분야의 통합으로 새로운 지식과 가치를 창출하며, 더 나은 정책을 수립하고 집행하는 과정을 마련하는 것이 핵심이라고 할 수 있다.

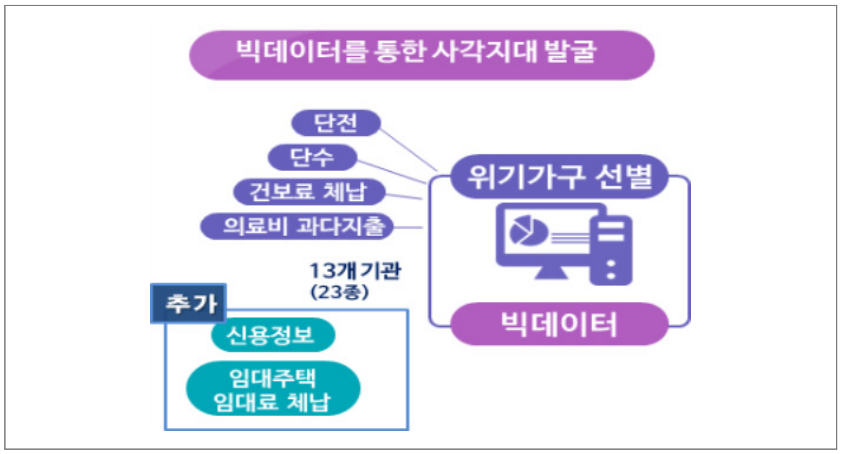
4차 산업혁명의 핵심 기술은 인공지능과 빅데이터로, 빅데이터를 활용한 기계학습(Machine Learning) 기반 자가 진화 알고리즘 성능의 지속적인 강화가 모든 분야의 중요한 원천 기술이라고 할 수 있다. 인공지능이란, 인간의 지능으로 가능한 인식, 판단, 추론, 문제 해결, 그 결과로서의 언어나 행동, 더 나아가 학습 기능과 같은 인간의 두뇌 작용을 컴퓨터가 할 수 있도록 실현하는 기술을 의미한다. 인공지능은 1950년대 후반 등장하여 시대적 상황과 여건에 따라 발전과 후퇴를 반복하였지만, 1990년대 후반부터 기계학습과 함께 꾸준히 발전해 왔으며, 최근 알파고 및 4차 산업혁명 이슈로 크게 주목받고 있다. 빅데이터는 기존의 데이터베이스 관리 도구로 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서, 대용량의 정형·비정형 데이터 집합과 이러한 데이터에서 가치를 추출하고 결과를 분석하는 기술도 포함하며, 빅데이터 분석의 중요성은 매우

빠르게 커지고 있다. 데이터로부터 생성되는 부가가치 측면에서 보면, 데이터를 분석해 새롭게 얻을 수 있는 지식 또는 부가가치의 양과 차이는 크지 않을 수 있지만, 국가 또는 기업 간 기술 격차의 감소로 인해 경쟁이 치열해지는 환경 속에서 빅데이터의 활용은 큰 차이를 가져오는 중요한 요인이 될 것이다.

기계학습(Machine Learning)은 AI 의 한 분야로 데이터를 바탕으로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이며, 이미지 처리, 영상인식, 음성인식, 인터넷 검색 등의 다양한 분야의 핵심 기술로 예측(Prediction)에 탁월한 성과를 나타낸다. 빅데이터 시대의 데이터 활용가치가 증대되면서, 다양한 기관으로부터 생산되는 사회보장 빅데이터를 활용하여 미래를 예측할 수 있는 보건복지 분야 활용 사례가 복지사각지대 발굴과 위기아동발굴시스템이라고 할 수 있다. 복지사각지대 발굴은 동절기 복지사각지대 발굴에 빅데이터를 활용하여 기계학습 기반 예측모형을 통해 고위험이 예상되는 11만4천명[2]을 선별하여 지자체가 전수조사를 실시하였으며, 1만7천명에게 필요한 서비스를 지원하였다. 이는 단전정보를 가지고 지자체가 전수조사를 하여 대상자에게 서비스를 지원하였던 경우보다 4배 정도 높은 성과를 보였다.

2) 2015년 12월 시점

〔그림 1-1〕 빅데이터를 통한 복지사각지대 발굴



자료: 보건복지부. (2017. 1. 9.). 보건복지부 2017 주요업무 계획 보도자료. p. 5.

또한, 아동의 권익 증진을 위해서 위기아동을 발굴할 수 있는 시스템(위기아동발굴시스템)을 구축(2017.7.~)하고 있는 상황이다. 위기아동발굴시스템은 장기결석, 건강검진 미실시 등의 빅데이터를 활용하여 학대 등 위기아동을 조기 발굴할 수 있는 시스템으로, 데이터 분석이 중요한 부분이며, 기계학습에 기반을 둔 예측모형을 적용할 수 있다.

〔그림 1-2〕 위기아동발굴시스템

자료: 보건복지부. (2017. 1. 9.). 보건복지부 2017 주요업무 계획 보도자료. p. 5.

캐나다 온타리오주에서도 대학, 연구원, 정부가 협력하여 아동 복지를 위해 행정데이터인 온타리오 아동학대 및 방임 데이터시스템(Ontario Child Abuse and Neglect Data System)을 활용하여 데이터를 분석한 사례가 있다(Fallon et al., 2017). 이를 통해 온타리오주의 아동 복지 서비스 제공에 관한 근거 기반 정책을 설계하고 아동 재학대에 중요한 요인을 확인할 수 있었다.

정책 관련 연구를 위한 행정데이터 중심의 사회보장 빅데이터 해외 활용 사례는 이웃 특성과 안전성에 관한 연구(O'Brien et al., 2015), 대학 졸업자들의 소득에 관한 연구(Britton et al., 2015), 교육 실험의 장기적인 영향에 관한 연구(Chetty et al., 2011a, b) 등이 있다.

이처럼 사회보장 빅데이터는 사회의 불평등 및 인간 행동을 연구하고 근거 기반의 사회 정책에 기여할 수 있는 강력한 자원이라고 할 수 있다. 여기에 기계학습 기법을 적극적으로 활용한다면, 앞서 언급한 복지사각지대 발굴 및 위기아동발굴시스템처럼 기계학습에 기반을 둔 사회보장 빅데이터 분석·연구를 통해 효과적인 정책 수립 및 집행으로 공공·행정 부문에서 효율성 증대가 가능하다는 것을 보여줄 수 있을 것이다.

2. 연구 목적

본 연구에서는 현재 국내에서 생산되고 있는 사회보장 빅데이터의 특성을 파악하고 기계학습(Machine Learning) 통계기법을 연구함으로써, 사회보장 빅데이터 분석에 적합한 기계학습 기반 예측모형을 설계하고 근거기반(evidence-based) 연구에 적용할 수 있는 방법론을 제시하고자 한다. 기계학습 기반 예측모형은 데이터의 활용 가능성을 높이고 다양한 분석을 가능하게 하여 새로운 가치를 도출하는 데 기여할 수 있다. 기계학습

은 방대한 데이터가 쌓이고 있는 사회정책 분야에 활용 가능하며 기계학습 기반 통계분석으로 예측 가능한 맞춤형 복지에 가까이 다가설 수 있게 할 것이다.

제2절 연구내용 및 방법

1. 연구 내용

이 연구에서는 사회보장 빅데이터와 기계학습 개념을 정의하고, 기계학습 기법이 사용된 다양한 활용 사례를 살펴보고, 최신 기계학습 기법이라고 할 수 있는 딥러닝(Deep Learning)에 대해 알아보고자 한다. 기계학습 기법에 대한 이론적 배경 소개와 각 방법론별 장·단점을 비교해 본다. 기계학습 기법을 적용한 예측모형을 평가할 수 있는 평가 방법을 살펴보고 한국복지패널을 이용하여 예측모형 시뮬레이션을 통해 여러 방법론을 비교·적용해본다. 마지막으로, 기계학습 기법 적용과 관련된 이슈로 알고리즘의 책무성과 데이터 주도 중심 정책을 다루어보고자 한다. 본 보고서는 모두 6개의 장으로 구성되어 있다.

2. 연구 방법

본 보고서 작성을 위해 국내·외 문헌연구, 해외사례연구 및 현지조사, 자료 분석, 국제 학회 자료집을 통한 해외동향 파악 등 다양한 방법이 활용되었다.

기계학습 기반 예측모형 방법론 비교 시뮬레이션에서 여러 통계적 기법을 고려하고 비교분석하기 위하여 R 프로그램을 사용하였다. 기계학습

통계기법으로는 Logistic, Elastic Net, Boosting, Support Vector Machine, Random Forest, Deep Learning을 이용하여 각 방법들에 대한 시뮬레이션을 실시하였다.

데이터 모의분석에서는 한국보건사회연구원의 한국복지패널을 사용하여 사회보장 빅데이터와 유사한 속성으로 데이터 구축 후 분석을 하였다.

제 2 장

기계학습(Machine Learning) 개념 및 활용 사례

2 기계학습(Machine Learning) 개념 및 활용 사례

제1절 사회보장 빅데이터 및 기계학습 개념 정의

1. 사회보장 빅데이터 개념

사회보장 빅데이터의 개념은 사회보장의 정의에 따라 사회보장이 포괄하는 영역의 범위와 관련된다. 빅데이터는 정형데이터와 비정형데이터로 나눌 수 있으며 행동 데이터(behaviour data), 이미지 데이터(image data), 소셜 데이터(social network data), 자연어 데이터(langage data) 등도 빅데이터의 범위에 포함된다. 기계학습 기법은 정형데이터뿐만 아니라 비정형데이터에서도 좋은 성능을 보여주고 있는데, 이 연구에서는 빅데이터의 범위를 정형데이터인 행정데이터로 한정하여 살펴보고자 한다. 따라서 사회보장 빅데이터는 "사회보장 분야에서 생산 및 구축된 행정데이터 또는 빅데이터"로 정의될 수 있다.

이는 사회보장기본법 제3조에 제시된 사회보장의 정의를 근거로 설정할 수 있는데, '사회보장'은 출산, 양육, 실업, 노령, 장애, 질병, 빈곤 및 사망 등의 사회적 위험으로부터 모든 국민을 보호하고 국민 삶의 질을 향상시키는 데 필요한 소득·서비스를 보장하는 사회보험, 공공부조, 사회서비스를 의미한다(제3조의 1). 여기서 '사회보험'과 '공공부조'는 전통적 사회적 위험에 대응하기 위한 대표적 복지정책 유형으로, 각각 "국민에게 발생하는 사회적 위험을 보험의 방식으로 대처함으로써 국민의 건강과 소득을 보장하는 제도(제3조의 2)"와 "국가와 지방자치단체의 책임하에

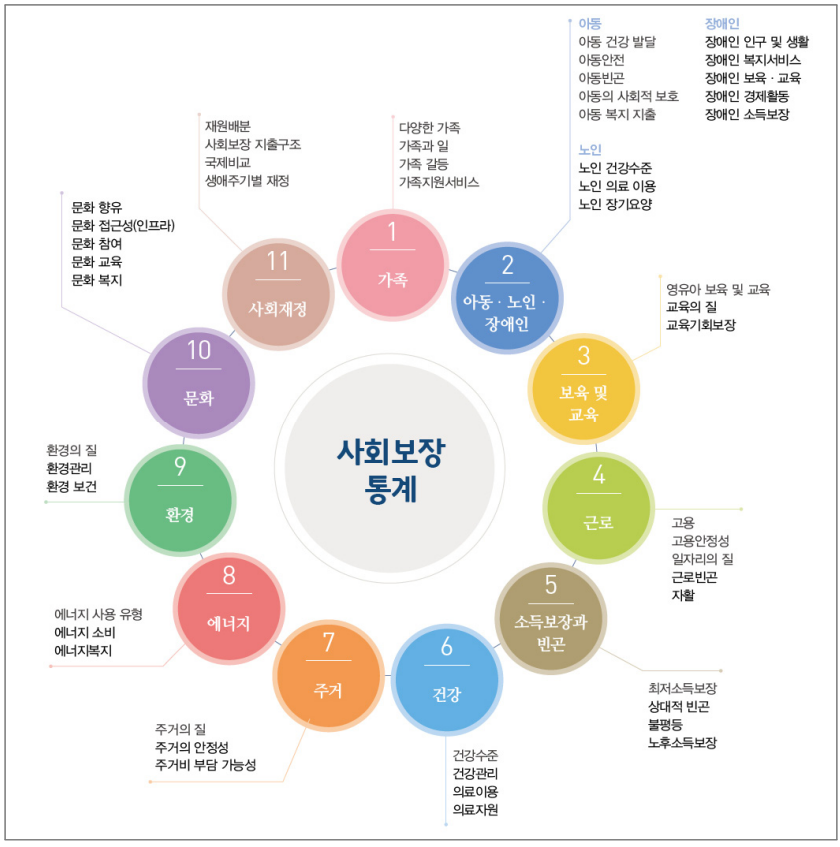
생활 유지 능력이 없거나 생활이 어려운 국민의 최저생활을 보장하고 자립을 지원하는 제도(제3조의 3)"를 의미한다. 이에 따라 사회보장 빅데이터의 개념에는 우선적으로 5가지 사회보험(국민연금, 건강보험, 고용보험, 산재보험, 장기요양보험) 제도의 운영과 관련된 행정데이터, 그리고 국민기초생활보장제도 등 다양한 공공부조제도 운영과 관련하여 구축된 행정데이터를 포함한다.

이러한 사회보험과 공공부조 외에 사회보장 빅데이터의 개념 및 범위를 설정하는 데 있어 중요한 개념은 사회서비스라고 할 수 있다. '사회서비스'란 국가·지방자치단체 및 민간 부문의 도움이 필요한 모든 국민에게 복지, 보건의료, 교육, 고용, 주거, 문화, 환경 등의 분야에서 인간다운 생활을 보장하고 상담 및 재활, 돌봄, 정보 제공, 관련 시설의 이용, 역량 개발, 사회참여 지원 등을 통해 국민의 삶의 질이 향상되도록 지원하는 제도를 의미한다(제3조의 4). 사회서비스 개념은 2013년 사회보장기본법의 시행에 따라 확장되었으며, 이에 따라 사회보장의 개념 및 범위는 전통적 복지정책의 영역뿐 아니라 국민의 삶의 질과 관련된 다양한 분야에서 시행되고 있는 정책들을 포괄하는 개념으로 정의될 수 있다.

이와 유사하게 사회보장의 개념과 범위가 설정 및 활용되고 있는 사례로 사회보장기본법 제32조에 근거해 매년 생산, 수집 및 공표되는 사회보장통계를 살펴볼 수 있다. 사회보장통계는 사회보장기본법에 명시된 사회보장 분야별 현황과 발전 단계를 파악할 수 있도록 표와 그래프를 활용하여 제시함으로써, 국민들로 하여금 사회보장 수준과 정책방향을 쉽게 이해하도록 만들고 정책수립을 위한 기초 자료로 사용할 수 있도록 제공함을 목적으로 한다. 사회보장통계는 지속적으로 근거 중심의 정책 평가 및 설계가 이루어져야 하며, 우리나라의 사회보장 실태와 분야별 현황을 진단할 수 있도록 영역별 주요 통계지표를 통합적으로 수집하여 관리 및

제공한다는 측면에서 중요한 의미를 지닌다. 사회보장 분야와 관련된 현황 및 문제 상황, 그리고 정책적 대응을 중심으로 사회보장통계의 범위를 설정하고 있으며, 향후 사회보장통계 구축 및 생산의 지속 가능성을 위하여 각 지표의 메타정보 및 유의점 등을 함께 제공하고 있으며, 이를 통해 국민들의 이해를 돕고 있다.

[그림 2-1] 사회보장통계 분야별 세부영역



자료: 이태진, 정해식, 최현수, 고제이, 김현경, 신정우 등. (2016). 통계로 보는 사회보장 2016. 세종: 보건복지부, 한국보건사회연구원.

이러한 사회보장통계에서는 사회보장의 개념 및 범위를 근간으로 생애주기와 기본적인 생활보장 영역에 따라서 사회보장통계의 하위 영역을 구분하고, 주요 정책분야별로 지표를 구성하여 체계적으로 제공하고 있다. “통계로 보는 사회보장 2016”에서는 사회보장통계의 영역으로 가족, 아동·노인·장애인, 보육·교육, 근로, 소득보장과 빈곤, 건강, 주거, 에너지, 환경, 문화, 사회재정 등 11개 분야로 분류하여 구성하고 있다. 이것은 최근 사회보장 분야의 복지정책 확대로 인해 보다 종합적으로 이러한 정책들을 살펴보고 향후 통합적인 관점에서 사회보장 실태와 정책을 평가하기 위한 목적을 지니고 있다.

이상에서 사회보장기본법에 기초하여 사회보장의 개념과 범위를 정의할 수 있는 것처럼, 행정데이터 또는 빅데이터를 의미하는 정보 및 이러한 정보가 구축 및 운영되는 시스템은 사회보장급여의 이용 · 제공 및 수급권자 발굴에 관한 법률(사회보장급여법) 제23조의 “사회보장정보”와 사회보장기본법 제37조에 근거한 “사회보장정보시스템(행복e음, 범정부 시스템)”으로 정의된다고 할 수 있다. 사회보장급여법 제23조(사회보장정보의 처리 등) 제1항에 “보건복지부 장관은 보장기관이 수급권자의 선정 및 급여관리 등에 관한 업무를 효율적으로 수행할 수 있도록 사회보장정보시스템을 통하여 다음의 각 호[3]에 해당하는 자료 또는 정보를 처리할 수 있다”고 규정하고 있으며, 이를 “사회보장정보”라고 지칭하고 있다.

3) 1. 근거 법령, 보장대상 및 내용, 예산 등 사회보장급여 현황에 관한 자료 또는 정보, 2. 제5조부터 제22조까지에 따른 상담, 신청, 조사 및 자격의 변동관리에 필요한 인적사항 · 소득 · 재산 등에 관한 자료 또는 정보, 3. 사회보장급여 수급이력에 관한 자료 또는 정보, 4. 제51조에 따라 보건복지부 장관이 위임 · 위탁받은 업무를 수행하는 데 필요한 자료 또는 정보, 5. 사회보장정보와 관련된 법령 등에 따른 상담, 신청(제25조 제3항에 따른 신청을 포함한다), 조사, 결정, 제공, 환수 등의 업무처리내역에 관한 자료 또는 정보, 6. 사회보장 관련 민간 법인 · 단체 · 시설의 사회보장급여 제공 현황 및 보조금 수급이력에 관한 자료 또는 정보, 7. 그 밖에 사회보장급여의 제공 · 관리 및 사회보장정보시스템 구축 · 운영에 필요한 정보로서 대통령령으로 정하는 자료 또는 정보.

요컨대, 사회보장 빅데이터의 개념과 범위는 “사회보장 분야(사회보장기본법 제3조)의 다양한 정책수립 및 시행과 관련하여 정책의 집행과정에서 사회보장정보시스템(사회보장기본법 제37조)과 각종 사회보험 제도 운영을 지원하는 정보시스템을 통해 구축된 사회보장정보(사회보장급여법 제23조 제1항)”로 설정할 수 있다.

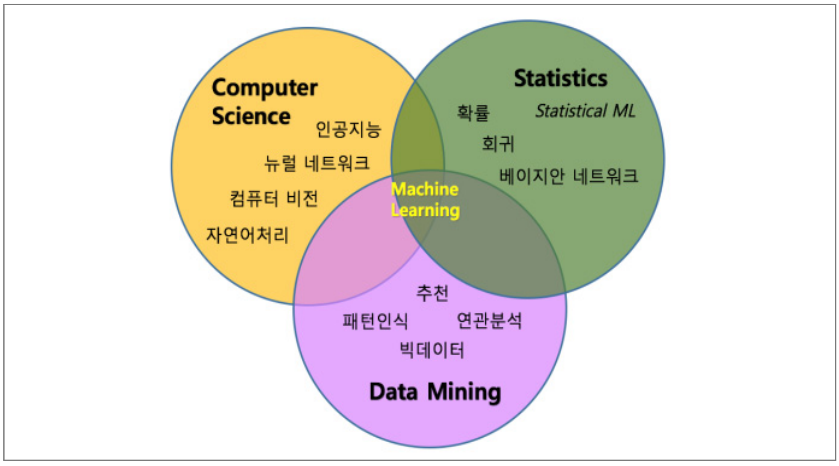
2. 기계학습 개념

기계학습 개념을 이해하기 위해서는 학습(learning)의 개념부터 이해하여야 한다. 실무적인 관점에서 학습의 정의(김의중, 2016, p. 77)는 표현(representation), 평가(evaluation), 최적화(optimization)의 합이라고 할 수 있다. 표현은 어떤 업무를 수행하는 에이전트가 입력값을 처리해서 어떻게 결과값을 만드는지를 결정하는 방법을 의미한다. 이는 모형이라고 설명할 수 있으며, 예를 들어 컴퓨터에 필기체를 인식하는 학습을 시킨다고 했을 때, 필기체 아라비아 숫자가 실제 어떤 숫자를 의미하는지를 인식하고 분류하는 논리모형이라고 할 수 있다. 평가는 에이전트가 얼마만큼 업무를 잘 수행했는지를 판정하는 방법을 의미한다. 이는 컴퓨터에 필기체를 인식하는 학습에서 필기체를 정확히 구분한 확률로 설명될 수 있다. 최적화는 평가에서 설정한 기준을 최적으로 만족하는 조건을 찾는 것이다. 최적화 과정이 끝나면 학습 모델에 사용된 가중치가 결정된다. 이는 학습이 완료되었다고 표현하며, 여러 가지 방법에 의해 학습이 완료된 후, 새로운 데이터에 대한 예측을 하는 것을 일반화(generalization)라고 한다.

기계학습은 데이터 마이닝과 혼용되기도 하는데(Domingos, 2012, pp.78-87), 기계학습은 컴퓨터가 사전에 프로그램되어 있지 않고 데이

터로부터 패턴을 학습하여 새로운 데이터에 대해 적절한 작업을 수행하는 일련의 알고리즘이나 처리 과정을 의미하는데 반면 데이터 마이닝은 주로 사람에게 어떤 지식(information or insight)을 제공하는 것을 목적으로 하고 있다. 기계학습은 컴퓨터 과학(Computer Science), 통계학(Statistics), 데이터 마이닝(Data Mining) 등의 분야와 밀접한 관련이 있는데, 통계학 분야에서는 기계학습을 통계적 학습(Statistical Learning)이라고도 한다.

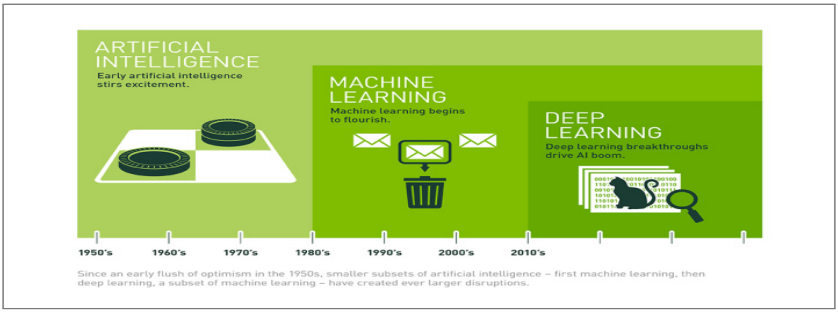
[그림 2-2] 기계학습 관련 분야



자료: Tensorflow blog. (2017a). 머신러닝이란. https://tensorflow.blog/%ed%95%b4%ec%bb%a4%ec%97%90%ea%b2%8c-%ec%a0%84%ed%95%b4%eb%93%a4%ec%9d%80-%eb%a8%b8%ec%8b%a0%eb%9f%ac%eb%8b%9d-1/에서 2017. 11. 29. 인출.

기계학습보다 큰 범위는 인공지능으로, 기계학습은 인공지능 범위 안에 포함된다. 요즘 화두가 되고 있는 딥러닝(Deep Learning)은 기계학습의 한 분야이긴 하지만, 기계학습과 따로 구분하여 기술하기도 한다.

[그림 2-3] 인공지능, 기계학습, 딥러닝 포함관계



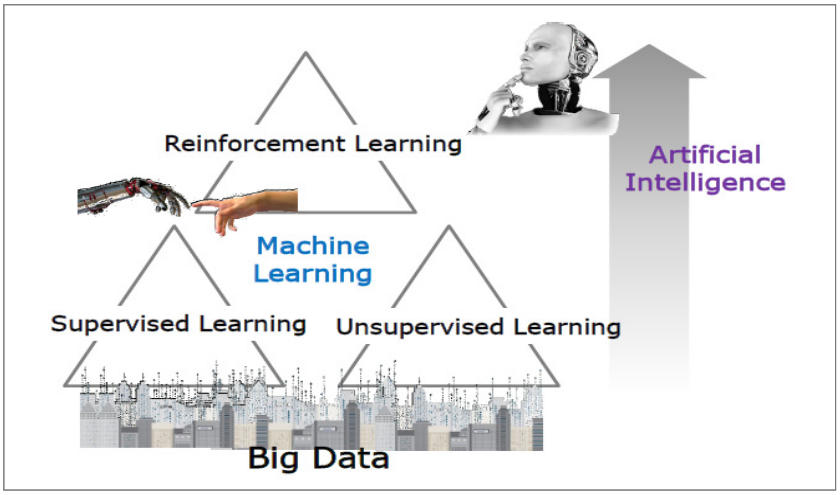
자료: Castrounis, A. (2016). Artificial Intelligence, Deep Learning, and Neural Networks, Explained. https://www.kdnuggets.com/2016/10/artificial-intelligence-deep-learning-neural-networks-explained.html에서 2017. 11. 29. 인출.

3. 기계학습 알고리즘 분류

기계학습은 학습 데이터(training data)에 레이블(label)이 있는 경우(ex. 종속변수 또는 반응변수)와 있지 않은 경우로 나누어 기계학습 모델을 학습시키는 방식을 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 설명한다. 레이블은 학습 데이터의 속성을 분석하고자 하는 관점에서 정의하는 것이다(김의중, 2016, p. 78). 지도학습에는 분류(classification)와 예측(prediction)모델이 있고, 비지도학습에는 군집(clustering)모델이 있다. 지도학습이나 비지도학습 외에 강화학습(reinforcement learning)은 지도학습에 포함된 모델로 분류하기도 하고 독립적인 기계학습의 한 종류로 구별하기도 한다. 강화학습은 알고리즘이 수행한 결과에 따라 수행 방식을 진화시켜 나가는 것이다. 데미스 하사비스(Demis Hassabis) 구글 딥마인드 대표는 알파고 2.0을 소개하며 "인간의 기보는 입력하지 않았다"라고 하였는데 이는 바둑의 규칙만 입력한 뒤, 스스로 바둑을 두며 이기는 수를 배워 나가는 것을 의미한다(중앙

일보, 2017). 2017년 10월 '네이처'에 소개된 알파고제로(AlphaGo Zero)가 강화학습 기반 알고리즘으로, 강화학습은 스스로 시행착오를 겪으며 학습하고 진화한다는 점이 지도학습이나 비지도학습과 큰 차이점이라고 할 수 있다. 강화학습은 시행착오 과정을 거쳐서 학습하고 진화하기 때문에 사람의 학습방식과 유사하며, 일부에서는 강화학습이 인공지능의 핵심이라고 이야기하기도 한다.

[그림 2-4] 기계학습 알고리즘 분류



자료: 최대우. (2017). 4차 산업혁명 시대의 빅데이터 그리고 AI. ICT Convergence Korea 2017 특강자료. 서울.

기계학습은 학습된 알고리즘으로 새로운 데이터를 처리하는 데 초점이 맞춰져 있어 예측력은 좋으나 기계학습으로 도출된 결과에 대해 해석하기 어려울 수 있다. 하지만 최근에는 해석하기 어려운 기계학습 방법론들도 설명이 가능할 수 있도록 시각화 툴 등을 사용하여 여러 가지 방법들을 개발 및 연구하고 있는 상황이다.

제2절 기계학습 기법 활용 사례

기계학습 기법들은 현재 산업 및 여러 학문 분야에서 다양하게 활용되고 있다. 기계학습, 특히 딥러닝이 활용되면서 기존에 해결되지 못하던 부분들이 해결될 뿐 아니라 성능적인 측면에서도 고무적인 발전을 가져왔다. 국내·외 유수 기업들인 구글, 페이스북, 마이크로소프트, 네이버, 삼성, LG, 아마존 등에서 기계학습 및 딥러닝의 기술을 기반으로 한 제품 및 소프트웨어들을 활발히 만들고 있는 모습들을 쉽게 찾아볼 수 있다. 이번 절에선 기계학습 및 딥러닝의 산업 및 학문 분야에서의 다양한 활용 사례들에 대해서 알아보고자 한다.

1. 기계학습 기법 활용 개관[4)]

기계학습이 가장 큰 강점을 보이는 분야는 패턴인식이라고 할 수 있는 이미지 및 영상, 음성 인식과 자연어처리이다. 딥러닝은 방대한 데이터를 분석해 이들의 차이점을 가려내고 유사한 것들을 분류하는 데 강점을 지니고 있다. 기계학습의 이 같은 강점이 어느 분야에 적용될 수 있는지를 먼저 살펴보고자 한다.

가. 얼굴을 포함한 물체 인식 기술

구글은 2012년 스탠퍼드 대학의 앤드루 응(Andrew Ng) 교수와 함께 진행한 딥러닝 연구 프로젝트인 구글 브레인을 통해 컴퓨터가 사진에서 고양이를 인식하도록 학습하는 데 성공했다. 구글 브레인은 분산 컴퓨터

4) 한국컨텐츠진흥원(2016), pp. 12-15를 참고하여 재구성.

시스템으로 1만 6000개 이상의 CPU 코어를 사용하여 딥러닝 알고리즘을 훈련시키고, 유튜브에서 1000만 건 이상의 고양이 동영상 중 고양이가 찍힌 사진을 추출한 뒤 딥러닝 시스템을 적용하였다. 그 결과 딥러닝 시스템은 다양한 이미지에서 고양이를 인식하는 데 성공했고, 이미지 인식 성공률은 기존의 인공지능 시스템보다 70% 높은 정확도를 보였다.

페이스북의 AI 연구소에서는 사람의 얼굴을 인식해 구분하는 딥페이스라는 이름의 기술을 개발해 2014년에 공개했다. 연구 논문에 따르면 이 기술을 이용한 얼굴인식 정확도는 97.25%로 인간의 평균 눈 수준(97.53%)에 가깝다.

나. 음성 인식과 자연어 처리

기계 학습에 기반한 음성 인식 기술과 자연어 처리 기술이 결합하면 인간의 말을 인식하고 이에 적절히 반응할 수 있는 인공지능 서비스 개발이 가능하다. 애플의 시리(Siri), 구글의 나우(Now), 삼성의 빅스비(Byxby)와 같은 인공지능 비서 서비스는 다양한 알고리즘의 기계학습 기술과 함께 음성 인식이 가능한 딥러닝 기술이 적용된 대표적인 사례이다. 인공지능 서비스 영역은 인터넷뿐만 아니라 가전제품에도 영향을 미치고 있는데, 딥러닝 기술이 탑재된 TV, 냉장고, 에어컨은 인간의 음성 및 주변 환경을 인지하고 사물과의 상호작용을 가능하게 한다..

다. 무인자율주행 자동차

기계학습을 활용한 영상인식 기술은 무인자율주행 차량 분야에도 사용되고 있다. 현재 테슬라, 아우디, 현대자동차 등 주요 자동차 회사뿐 아니라 구글, 네이버, 심지어 엔비디아까지 자율주행차량 기술을 확보하기 위

해서 다양한 기계학습 기술들을 활용하고 있다.

라. 계속 확대되는 기계학습 활용 분야

IBM의 인공지능 슈퍼 컴퓨터 왓슨(Watson)은 인공지능 플랫폼 중에서 가장 널리 알려져 있다. 왓슨은 자연어처리 분석 기술과 기계 학습 기술을 활용하여 정형화되어 있지 않은 데이터들 가운데서 의미와 직관적 정보를 도출할 수 있는 인공지능 기술 플랫폼이다. 왓슨은 빅데이터와 인공지능이 결합한 모형으로 다양한 용도로 활용되기 시작하였는데, 최근에는 미국 건강보험업체 웰포인트의 컨설턴트로 고용되어 의료서비스 관련 데이터, 질병 데이터, 환자 데이터를 분석하여 가장 적합한 답을 제시해주고 있다.

또한 화제의 중심에 있었던 구글의 알파고는 기계학습을 바둑에 적용하였다. 알파고에 적용되는 알고리즘은 딥러닝 기반 기계 학습과 트리 탐색 기법을 활용하고 있다. 구글은 기계학습 기술을 발전시켜 구글의 서비스와 다양한 분야에 적용하는 것을 목표로 하고 있다.

향후 기계학습의 적용분야는 지속적으로 확대될 것으로 예상된다. 기계학습 기술은 민간 기업들의 제품 개발뿐만 아니라 공공영역에서의 여러 사회적 문제를 해결하는 데에도 도움을 줄 수 있을 것이다.

2. 기계학습 기법의 해외 활용 사례

글로벌 IT기업들을 중심으로 기계학습 개발에 적극적으로 나서고 있다. 최근 기계학습 연구는 막대한 자금과 최고 수준의 인력을 바탕으로, 학습을 위한 빅데이터 확보가 가능한 기업 주도로 연구개발이 활발히 진

행 중이다. 특히 인공지능 연구 분야의 기계학습과 딥러닝 기술의 개발이 가속화되고 있다. 학습을 위한 막대한 양의 빅데이터 계산처리에 GPU를 이용한 병렬 연산 기법을 활용하여 성능이 크게 향상되었다. 글로벌 기업들은 앞다투어 Yann LeCun, Jeoffrey Hinton, Andrew Ng, Demis Hassabis 등 기계학습 관련 고급인력 확보를 위한 인수합병이 급증하고 있는 추세이다. 이러한 기계학습 관련 고급인력들은 IBM, 구글 등 기업의 부가가치를 창출하는데 큰 기여를 하고 있다.

IBM의 인공지능 시스템 왓슨(Watson)은 딥러닝 기법을 추가하여, 2011년 미국의 퀴즈쇼 '제퍼디'에서 승리하였다. 왓슨은 자연어처리, 기계학습 기술을 이용하여 개방형 질문에 대답할 수 있고, 퀴즈 문제를 일상 언어 문장으로 받아들여 인공지능을 바탕으로 고도의 지능문제를 분석해 답을 찾아냈다. 특히 의료분야에서도 왓슨으로 상업화를 진행하고 있는데, 전 세계적 지명도를 가진 암 전문 병원 의사들을 도와 암 여부를 판별하고 치료방법을 제안하는 임무를 수행 중이다. Texas MD Anderson 암 센터, 존슨앤존슨(Johnson & Johnson), 뉴욕 유전자센터(New York Genome Center), 베일러 의과대학(Baylor College of Medicine) 등 헬스케어기업 및 기관들과 제휴를 늘려 가고 있는 중이다. IBM에 따르면 의사의 수작업에 의존할 경우 최장 5~10개월 걸리던 작업을, 왓슨의 경우는 단 몇 분 만에 수행할 수 있으며, 시간이 지날수록 왓슨의 학습과정에 사람이 개입할 여지도 크게 줄어들 것으로 전망하고 있다. 또한, 번역 기능 및 음성을 텍스트로 혹은 텍스트를 음성으로 변환하는 기능을 일반 개발자에게 공개, 이를 활용한 서비스를 구축 가능케 하고 있다. 최근 의약, 교육, 출판에 관한 조사를 목적으로 하는 서비스인 Watson Discovery Advisor를 발표하며, 연구자들에게 데이터와 관련 내용을 보다 신속하게 얻을 수 있도록 하였다.

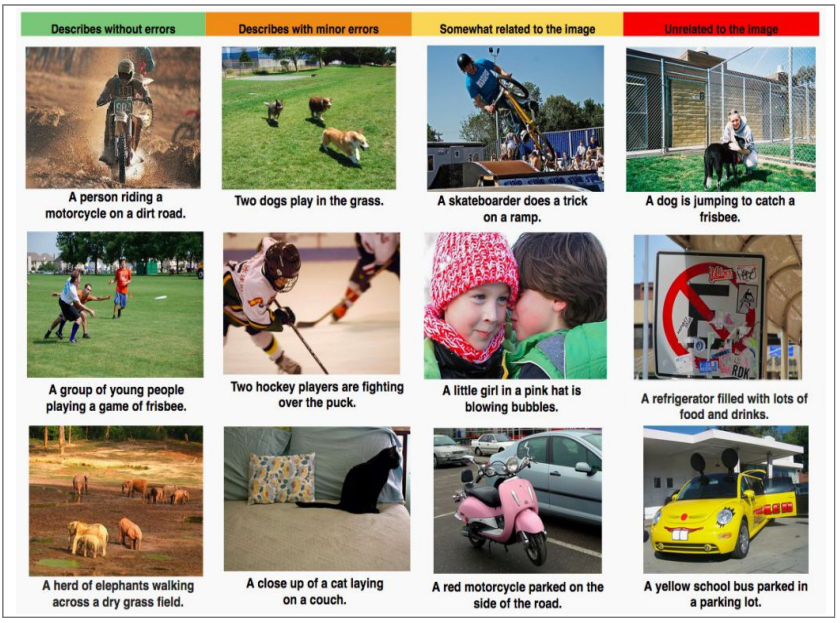
[그림 2-5] 2011년 퀴즈쇼 '제퍼디'에 참가한 IBM의 왓슨

자료: 김정형. (2013). 컴퓨터와 인간의 게임 대결, 어디까지 왔나. Premium Chosun. http://premium.chosun.com/site/data/html_dir/2013/12/26/2013122602309.html에서 2017. 11. 29. 인출.

구글은 2011년 브레인 프로젝트(Brain Project)를 수립하고, 기계학습 관련 연구개발을 지휘했던 앤드루 응 교수가 바이두(Baidu)로 이직하면서 타격을 받았으나, 기계학습의 대가인 레이 커즈와일(Ray Kuzweil)을 기술책임자로 영입하고, 강화학습(Reinforcement Learning)에 전문성을 보유하고 있는 영국의 스타트업인 딥마인드(DeepMind)를 인수하여, 기계학습 분야에서 세계 최고 수준의 기술을 보유한 것으로 평가받고 있다. 구글은 딥러닝에 필요한 많은 양의 학습 데이터를 수집하기 위하여, 2011년부터 약관을 수정하여, 사용자들로부터 막대한 양의 데이터를 수집하여 다양한 서비스를 위한 학습데이터로 활용하고 있다. 스마트폰 사용자를 위해 이메일 내용을 분석하고, 사용자의 모든 동작을 파악하며, 묻기 전에 사용자의 의도를 파악해서 검색 후 결과를 보여주는 사이버 도우미(Cybernetic Friend)를 개발하는 것을 목표로 삼고 있다. 또한 컴퓨터 비전 분야에서도 완벽한 문장을 사용하여 사진 속 장면을 정확하

게 묘사할 수 있는 소프트웨어를 개발하였다. 이와 같은 기계학습 기술을 바탕으로 자율주행 자동차와 로봇 등에서 강력한 경쟁우위를 확보하여 기계학습 분야에서 두각을 나타내고 있다.

[그림 2-6] 구글의 사진 묘사 기술



자료: Andrej, K. & Li, F. F. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. CVPR 2015 paper.

구글은 2015년 11월 오픈소스 기반의 인공신경망 알고리즘인 텐서플로(Tensorflow)를 공개하였다. 구글의 에릭 슈미트 회장은 "구글이 기계학습 시스템의 소스코드를 공개해 표준을 주도하겠다"라고 밝히며 기계학습 시스템 표준의 필요성을 공개적으로 역설하였다. 이에 맞춰, 실제로 구글 내부에서 사용하는 인공신경망 알고리즘을 텐서플로라는 이름으로 오픈소스화 하여 공개하였다. 텐서플로는 텐서(Tensor) + 데이터 플로

(Data Flow)의 합성어로, 텐서는 수학의 다중 선형함수를 말하며 모든 데이터를 수치화된 벡터값으로 표현한 다차원 배열을 의미하며, 데이터 플로는 하나의 작업을 수행하기 위하여 실행되는 각각의 세부작업들 사이에서 자료가 입출력되는 것으로, 복수의 인공신경망 연산 작업이 병렬처리됨을 의미한다.

텐서플로의 특징으로는(Tensorflow, 2017b),

1) 범용성과 속도: 텐서플로는 그래디언트(Gradient)기반의 모든 기계학습 알고리즘과 함께 사용할 수 있어서 그 응용범위가 매우 넓다. 속도를 위해서 C++로 작성되었으며, 개발자는 하드웨어에 대한 별도의 지식 없어도 개발 가능하다. 텐서플로는 초창기 구글의 인공신경망 시스템인 디스트빌리프(DistBelief)보다 최대 5배나 빠른 속도로 신경망을 구축하고 학습할 수 있는 것으로 나타났다.
2) 용이성과 확장성: 텐서플로는 파이선(Python) 및 C/C++ 를 지원하여 사용의 용이성 측면에서 강점을 갖는다. 파이선 인터페이스를 통해 도구, 예제, 튜토리얼과 같은 독립형 라이브러리를 사용할 수 있고, 단일 API를 사용하여 다양한 장치(데스크톱, 서버, 모바일 등)에서 하나 이상의 CPU 및 GPU에게 연산을 수행시킬 수 있는 확장성을 제공한다.
3) 오픈소스: 텐서플로는 현재 모든 사람들에게 공개되어 있고, 아파치 2.0 라이선스 조건하에서 누구나 사용 가능하다. 현재 공개된 버전은 단일 컴퓨터에서만 동작할 수 있으나, 이를 확장한 기능이 조만간 출시될 예정이다.

2015년 말, 텐서플로의 공개와 함께 이에 기반을 둔 구글포토(Google Photos)라는 서비스를 함께 발표하였는데, 구글포토는 사용자들의 사진들을 자동으로 분류하고 명시적인 태깅 정보가 없어도 검색을 가능하게

해주는 등 사진의 맥락을 이해하는 기계학습 기술의 상용화를 압도적으로 주도하고 있다. 구글포토는 출시 6개월 만에 1억 명 이상의 사용자를 확보하였으며, 이들이 500억 장이 넘는 사진을 등록하였다.

2015년 구글 딥마인드는 기계학습 바둑프로그램 알파고(AlphaGo)를 선보였다. 알파고에는 CNN(Convolutional Neural Network)을 이용한 딥러닝 기술과 바둑프로그램으로 널리 쓰이는 Monte Carlo Tree Search 알고리즘이 사용되었다. 발표 당시 프로 2~5단 수준이었으나, 현재는 프로 9단 수준으로 평가받고 있다. 유럽 바둑 챔피언이자 중국 프로 바둑기사인 판후이(2단)와 다섯 차례 대국을 벌여 모두 승리했으며, 2016년 이세돌(9단)과의 대국에서도 4승 1패로 승리하여 세계를 놀라게 했다.

〔그림 2-7〕 알파고와 이세돌 대국

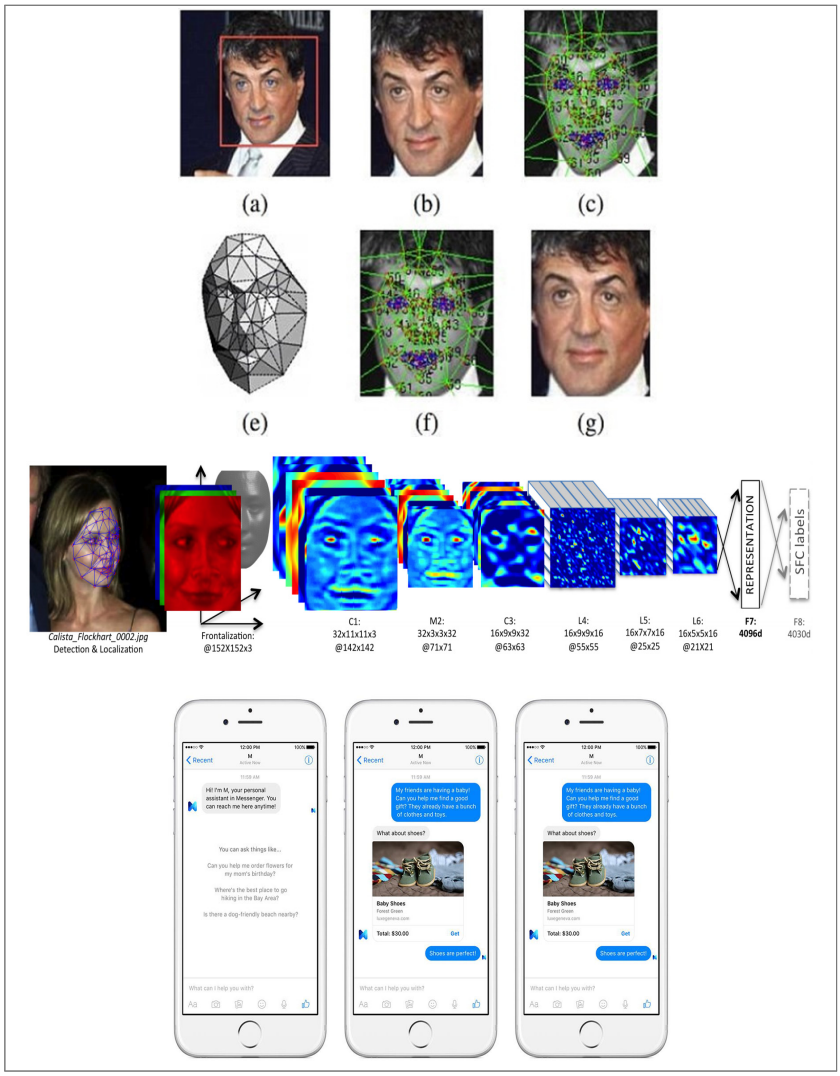
자료: 중앙일보. (2017. 10. 19.). 알파고 업그레이드...더 이상 인간의 기보 입력 안 한다. http://news.joins.com/article/22026687에서 2017. 11. 29. 인출.

페이스북(Facebook)의 기계학습 기술 또한 딥러닝 기반의 이미지 인식연구에 많은 투자를 해오고 있다. 사진 속의 사물과 인물 구별, 뉴스내용의 필터링, 자동 텍스트 번역, 음성인식 등을 수행하는 기계학습 기술을 보유하고 있다. 특히 얼굴인식 알고리즘 딥페이스(DeepFace)는 사람

의 얼굴을 97.25%의 정확도로 알아낼 수 있으며, 인간의 평균 정확도 97.35%에 근접한 것으로, 실제 페이스북 내에서 활용되고 있다. 딥페이스는 각기 다른 앵글로 찍은 두 개의 얼굴 사진을 3D모델을 이용해 회전시켜 정면을 바라보도록 조정한 다음, 딥러닝 기법을 이용해 얼굴의 각각의 구성요소에 매겨진 값이 두 사진 사이에서 얼마나 서로 일치하는지를 판단하여 사람의 얼굴을 인식하게 된다. 이 역시, 페이스북이 저장하고 있는 수억 장 이상의 얼굴 사진과 사진에 붙어있는 태그 정보가 딥러닝의 학습데이터로 활용되어 매우 높은 정확도를 확보할 수 있었다.

페이스북은 가상비서 서비스 'M'을 출시하였다. M은 메신저 앱을 활용하는 텍스트 기반의 서비스로 식당 예약, 선물 추천, 휴가지 추천 등의 요청을 수행할 수 있는 인공지능과 인간지능을 혼합한 서비스이다. 또한 2015년 페이스북은 구글과 더불어 GPU에 특화된 오픈소스의 딥러닝 모듈인 토치(Torch) 및 인공지능 분석 서버 빅서(Big Sur)를 공개하는 등 표준화에 노력하고 있다. 빅서는 기계학습 데이터를 학습할 때 사용되는 서버로, 데이터 처리 속도를 높인 것이 특징이다.

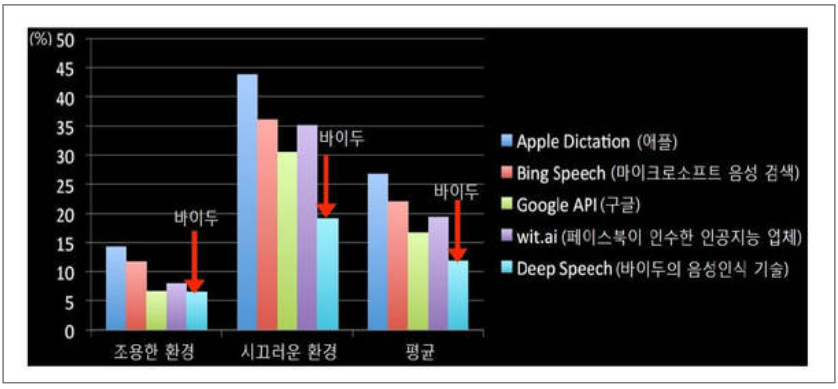
[그림 2-8] 페이스북의 기계학습 적용 사례



자료: 허핑턴포스트. (2014. 3. 21.). 페이스북, 얼굴인식 기능 '딥 페이스' 개발. http://www.huffingtonpost.kr/2014/03/21/story_n_4996658.html에서 2017. 11. 29. 인출.

바이두(Baidu)는 2013년 베이징과 실리콘밸리에 딥러닝 연구소를 설립하고 앤드루 응 교수 총괄 아래 이미지인식, 음성인식, 행동인식을 위해 딥러닝 기반의 Deep Image, Deep Speech, Deep Behavior 기술을 개발하고 있다.

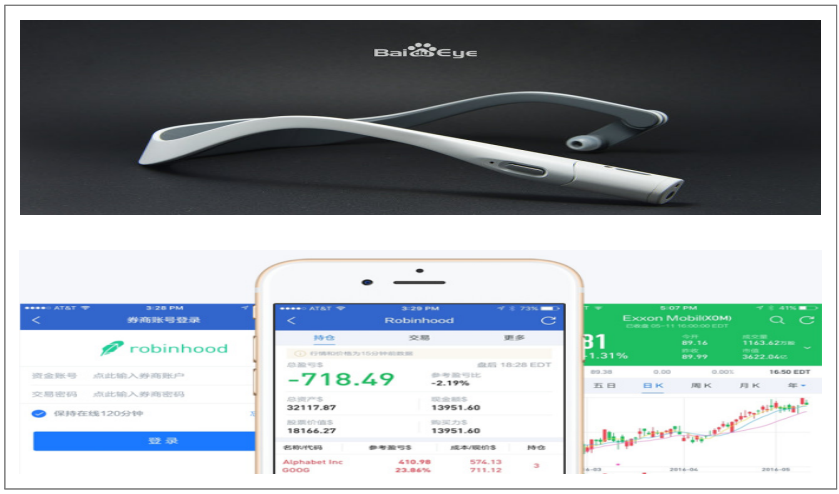
[그림 2-9] 주요 음성인식 기술들의 오인식률 비교



자료: 유진상. (2015. 8. 6.). 이제는 인공지능시대…ICT 기업들 경쟁 점화. 조선비즈. http://it.chosun.com/news/article.html?no=2805202에서 2017. 11. 29. 인출.

사용자에게 '핸즈프리(hands-free)' 검색 경험을 제공하는 것을 목표로, 머리에 착용하는 웨어러블 형태의 바이두아이(BaiduEye)를 발표, 구글 글라스와 마찬가지로 물체를 스캔하고 인식하는 기능을 지원한다고 밝혔다. 또한 바이두는 뉴스와 주식시장, 검색엔진에 기계학습 기술을 적용하여 주가와 테마주를 예측하는 스톡마스터(StockMaster)를 출시하여 금융분야에 기계학습을 접목하는 등 기계학습 연구에 집중하고 있다. 자사 실리콘밸리 연구소에서 개발한 WARP-CTC는 컴퓨터가 사람의 말을 인식하기 위한 기계학습 소프트웨어이며, 오픈 소스로 공개하였다. 이를 통해 딥스피치(Deep Speech)2 엔진을 개발한 것으로 알려졌다.

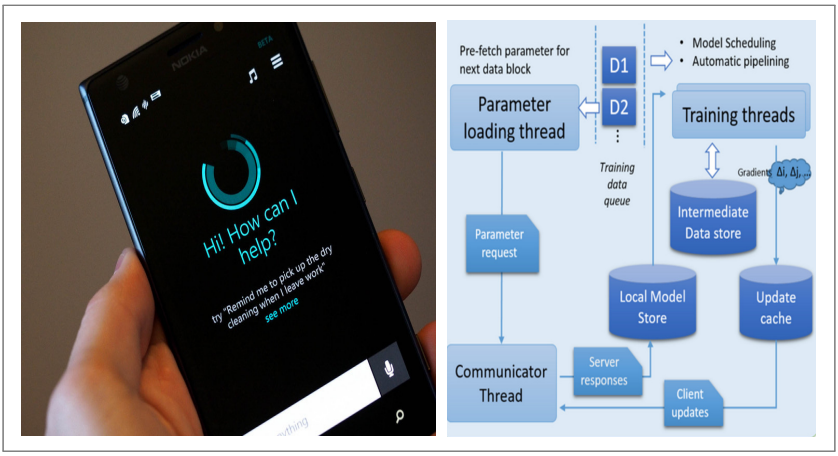
[그림 2-10] 바이두아이의 Stockmaster



자료: Constine, J. (2016). Robinhood lets china trade US stocks free through Baidu's finance app. https://techcrunch.com/2016/06/01/robinhood-china/에서 2017. 11. 29. 인출.

마이크로소프트(MS)는 인공지능 개인비서 코타나(Cortana)를 출시하였다. 코타나는 구글의 나우와 애플의 시리와 마찬가지로 사람과 주고받는 대화를 통해 상호작용하는 기계학습 기반 알고리즘으로, 코타나를 사용하면 할수록 더 똑똑해지도록 디자인 되었다. 더불어 스카이프(Skype)에서 실시간 언어 번역, 이미지 내의 물체를 인식하는 화상인식 기술인 아담 프로젝트를 진행 중이다. 또한, 마이크로소프트는 구글에 이어 2015년에 기계학습 툴킷인 DMTK(Distributed Machine Learning Toolkit)를 오픈소스 프로젝트 공유사이트인 깃허브(GitHub)에 올리며 공개하였다. DMTK는 서버-클라이언트 구조를 위한 C++ 기반의 SDK (Software Development Kit)이고, 한 번에 여러 노드 상에서 모델을 훈련할 수 있도록 하였다. 마이크로소프트는 다른 기계학습 알고리즘도 공개할 계획인 것으로 알려져 있다.

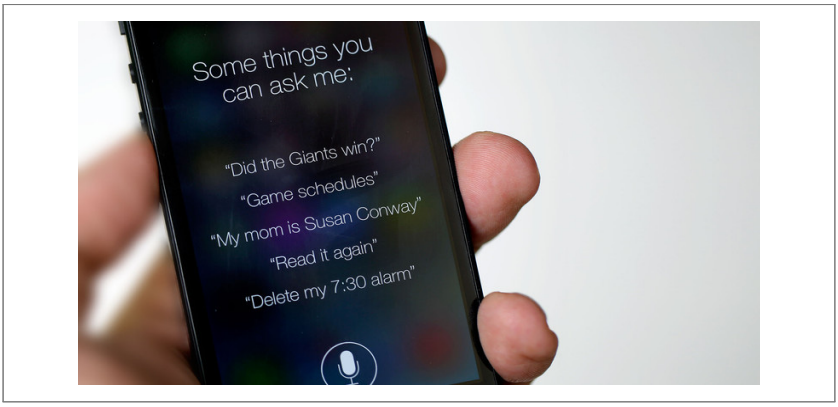
[그림 2-11] 코타나, DMTK 프레임워크



주: 코타나(좌), DMTK 프레임워크(우), 이미지.
자료: Science X Network. (2015). http://phys.org에서 2017. 11. 29. 인출.

애플은 2011년 출시한 음성인식 시리(Siri)의 인공지능을 강화하는 기술을 개발하고 있다. 음성인식 기반의 인터페이스를 통해 사물인터넷(IoT)과 자율주행 자동차에 활용될 것으로 예상된다.

[그림 2-12] 애플 시리



자료: Apple. (2011). https://www.apple.com/kr/ios/siri/에서 2017. 11. 29. 인출.

아마존에서는 신속한 배송과 정확한 물류관리를 목표로 기계학습 기술을 도입하였다. 아마존은 기계학습 알고리즘 기반으로 고객의 구매 패턴을 분석하여 이용자가 관심 있을 만한 상품을 추천해 왔다. 구매 추천 서비스는 매우 효과적이어서 아마존 매출의 3분의 1 이상이 구매 추천 서비스에 의해 발생한다(김인중, 2016, p. 31.). 더 나아가 2014년에는 고객이 주문하기도 전에 배송을 시작하는 기술을 실용화 하였으며, 관련 특허도 확보하였다. 기계학습을 이용해 고객이 구매할 것으로 예상되는 물품을 미리 포장해서 고객과 가까운 물류창고에 옮겨 놓음으로써 배달 시간과 물류비용을 절감하였다. 사람의 말투와 억양, 문맥을 파악하는 클라우드 기반 음성인식 인공지능 알렉사(Alexa)를 출시하여 대중의 주목을 받고 있다. 저가형 에코 닷(Echo Dot 화면이 장착된 에코 쇼 등 점점 더 다양한 제품들을 선보이고 있다. 특히 알렉사 API를 공개함으로써 다른 회사들이 음성인식 기술과 알렉사의 인공지능을 이용한 제품을 쉽게 만들어 생태계를 구축하기 위해 힘쓰고 있다. 더불어 드론 배송과 로봇에 의한 물류창고 관리, 예측배송 시스템을 통한 자동주문 배송 시스템 개발에 박차를 가하고 있다. 아마존 웹서비스(AWS, Amazon Web Services)에서는 클라우드 서비스에 기계학습 기능을 추가해 제공하고 있다. 클라우드 이용자는 아마존이 보유한 기계학습 알고리즘을 활용하여 부정거래탐지, 요구예측, 맞춤형 콘텐츠, 사용자행동 예측, 소셜미디어 현황 분석, 텍스트 분석 등 인공지능을 요구하는 애플리케이션을 개발하고 구축할 수 있다.

[그림 2-13] 아마존 알렉사와 드론 배송

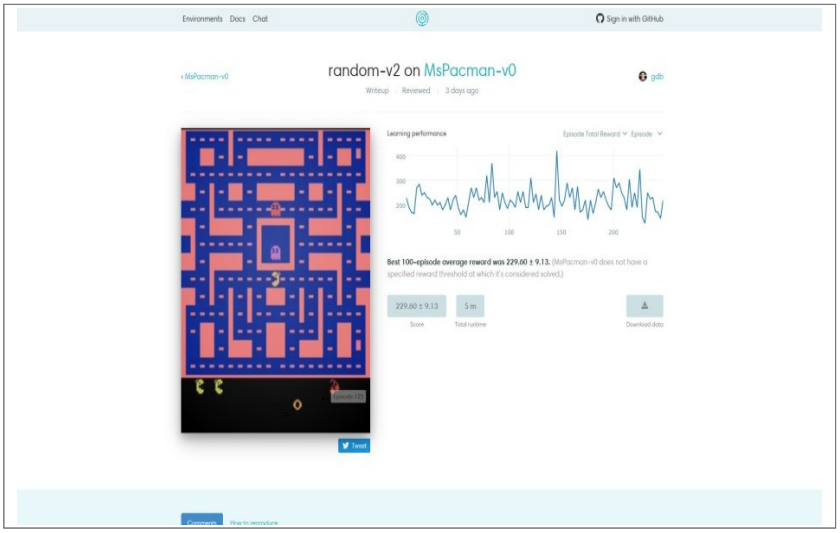
자료: Amazon. (2017). http://amazon.com/ 제품 설명 화면에서 2017. 11. 29. 인출.

야후는 자사의 오픈소스기반 딥러닝 소프트웨어 '카페온스파크(CaffeOnSpark)'를 공개했다. 카페온스파크는 아파치 스파크 프레임워크에 기반을 두고 있으며, 이미지 분석에 활용해왔다. 관련 업계에서는 이 기술이 공개됨으로써 음성인식 및 사진 또는 비디오의 내용을 식별할 수 있는 딥러닝 기술 개발에 활용할 수 있을 것으로 기대하고 있다. 또한 연구단체들이 대규모 데이터 기반 기계학습을 연구할 수 있도록 기계학습 예제 데이터를 공개했는데. 이는 2000만 사용자가 2015년 2월부터 5월까지 야후 뉴스피드에 보낸 정보로, 약 1.5TB 규모이다.

OpenAI는 인공지능을 사람 수준으로 향상시키는 기술 개발 목적으로, 테슬라 모터스 및 스페이스엑스(SpaceX)의 CEO 엘론 머스크(Elon Musk)와 유명 스타트업 액셀러레이터 와이콤비네이터(Y Combinator)의 CEO 샘 앨트먼(Sam Altman)에 의해 2015년 12월에 설립되었다. 공식 블로그에 따르면, 인공지능 기술을 개발하는 비영리 기업으로, 특정 사업이나 기타 목적에 얽매이지 않고 모든 연구 결과를 공개할 계획이라고 밝혔다. 리서치 디렉터로는 구글 출신 일리야 서츠케버(ilya Sutskever)

가, CTO는 인터넷 결제기업 스트라이프(Stripe) 출신의 그렉 브로크만(Greg Brockman), 고문에는 퍼스널 컴퓨터 및 객체 지향 프로그래밍 창시자인 앨런 케이(Alan Kay)가 참여하고 있다. OpenAI의 스폰서로는 페이팔의 공동 창립자 피터 시엘(Peter Thiel), 링크드인 공동 창업자 리드 호프먼(Reid Hoffman), 와이콤비네이터의 파트너 제시카 리빙스턴(Jessica Livingston), 그리고 아마존웹서비스(AWS), 인포시스, 와이시 리서치(YC Research) 등이 합류하였다. 소속 연구자는 자신의 성과를 논문, 블로그, 코드 등의 형태로 공개하고, 지식재산권은 전 세계와 공유할 예정이다. 다수의 조직 및 기업과 협력해 신기술을 개발할 것으로 기대된다. 2016년 4월 OpenAI는 강화학습(reinforcement learning) 알고리즘 개발 및 성능비교를 위한 툴킷 OpenAI Gym을 발표하여, 표준화를 위한 포석을 마련하고 있다.

〔그림 2-14〕 OpenAI Gym 활용 예제



자료: Intel Nervana. (2017). http://intelnervana.co/에서 2017. 11. 29. 인출.

테슬라 모터(Tesla Motors)의 엘론 머스크는 인공지능의 부정적 효과를 피력하면서도, 자율주행 자동차에 들어가는 기계학습 기술에 대한 연구개발을 통해, 2018년에는 부분적으로 자동화된 자율주행이 가능한 자동차가 생산될 수 있을 것으로 전망하였다.

[그림 2-15] 테슬라 자동차의 자율주행시스템(Autopilot)

자료: Heisler, Y. (2016. 7. 7.).
http://bgr.com/2016/07/07/tesla-autopilot-software-model-x-crash-nhtsa/ 에서 2017. 11. 29. 인출.

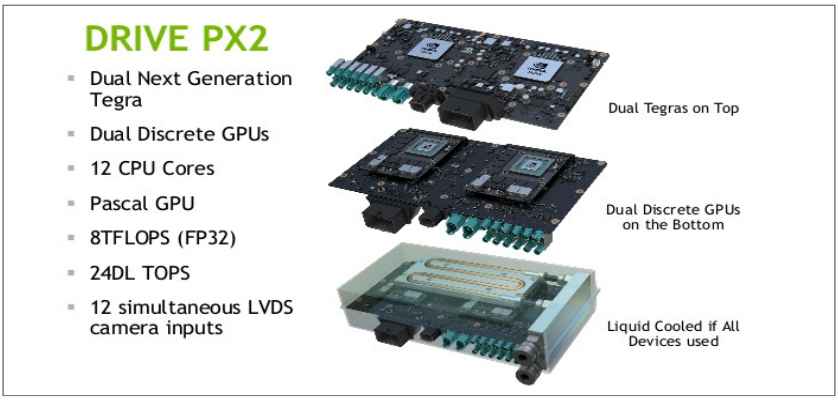
모빌아이(Mobileye)는 ADAS(Advanced Driver Assistance System)를 개발하여 상용화하였다. 컴퓨터비전 기술을 이용하여 차량 출동 가능성을 예측하고, 응급상황에서 자동차를 자동으로 정지시키는 시스템을 작동시키는 기능을 개발하였다.

컨티넨탈(Continental)은 프랑크푸르트에서 실시한 국제 모터쇼에서 기계학습과 관련한 연구개발 계획을 발표하였으며, IBM과 협력하여 자동차를 이용하는 고객이 운전에 보다 집중할 수 있도록 자연어처리과정 기술을 개발할 것이라고 언급하였다. 센서 기술이나 자동차 응용 애플리

케이션, 스마트폰 연동 기술 등과 관련된 연구도 지속적으로 진행 중이다. 이와 관련하여 2015년에 국토교통부 주도로 자율주행자동차 상용화 지원 방안을 마련하여 자율주행 자동차가 우리나라에 잘 정착될 수 있도록 지원하고 있다.

그래픽카드 제조사인 엔비디아는 HW 제조사 중 하나로 딥러닝 기술 향상에 큰 기여를 하고 있다. 최근에는 GPU 기반 차량용 딥러닝 시스템인 드라이브도 개발했다. 엔비디아는 CUDA라는 병렬 컴퓨팅 아키텍처를 개발하여 GPGPU 기술을 이용하여 컴퓨터 성능을 향상시켰다. 딥러닝은 CUDA의 성능을 활용하기에 매우 적합하기 때문에 그로 인해 엔비디아는 고성능 그래픽 카드의 판매를 크게 늘릴 수 있었다. 2016년 발표된 최신 시스템인 드라이브 PX2는 슈퍼컴퓨터에 필적할 성능과 신경망 네트워크 기반의 딥러닝 기능을 갖춰 자율주행 차량이 주위 환경을 관찰하고 분석하는 데 활용될 수 있다(한국컨텐츠진흥원, 2016, pp. 12-15). 드라이브 PX2는 자동차에 탑재된 센서, 카메라, 라이다, 레이더, 초음파 장비 등이 측정한 데이터를 딥러닝 연산으로 처리한다. 이 시스템은 카메라를 통해 촬영된 영상과 각종 센서와 레이더로 측정된 거리 정보 등을 종합해 영상에서 사람과 자동차를 판별해내는 것이다. 드라이브 PX2의 슈퍼컴퓨터급 성능은 도로에 떨어진 쓰레기나 장애물을 인식하고 강한 비나 눈이 내리는 혹한의 날씨에 자율주행 차량의 영상 인식 알고리즘을 처리하는 데 강점을 발휘할 것으로 전망된다.

[그림 2-16] 엔비디아 드라이브 PX2 구조



자료: NVIDIA. (2017). http://kr.nvidia.com/object/drive-px-kr.html에서 2017. 11. 29. 인출.

에이아이큐어(AiCure)는 미국 국립 보건연구소(NIH)가 출연한 기계학습 기반 생명과학 기업으로 약 복용 여부를 추적하고 환자의 행동 패턴을 분석하는 데 기계학습 기술을 활용하고 있다. 기계학습 기술이 적용된 플랫폼을 스마트폰과 연동시켜 치료약과 진료성과 사이의 관계를 파악하며, 실시간으로 환자의 처방약 복용 준수 여부를 확인하고 복용량을 지키지 않거나 스마트폰 앱을 사용하지 않을 경우 자동적으로 신호를 전송한다.

인터넷 마켓 플레이스 이베이(ebay)는 딥러닝 기법을 적용하여 판매자의 포스팅 이미지를 제품별로 분류하여 판매자가 태그를 붙이지 않은 경우에도 자동으로 제품을 분류할 수 있도록 하였다.

LA타임스, 포브스, AP, 영국 가디언 등은 내러티브 사이언스(Narrative Science)의 기사 작성 알고리즘을 이용해 기상예보, 금융시장 및 기업재무, 스포츠 등 데이터에 기반하여 사실을 전달하는 스트레이트성 기사를 송출 하고 있다. 또한 보안, 광고, 금융, 신약개발, 범죄예방 등 전방위적으로 기계학습을 활용한 다양한 비즈니스 기회가 늘어나고 있다.

3. 기계학습 기법의 국내 활용 사례

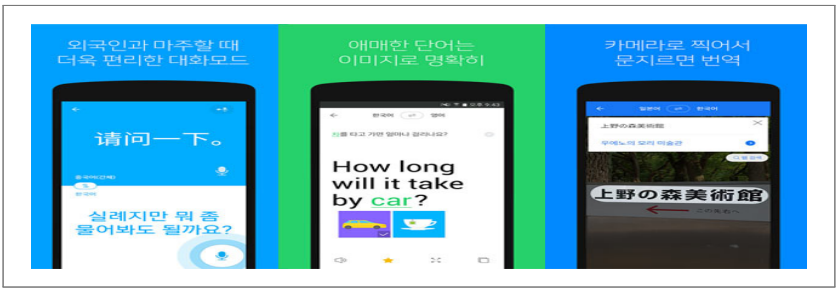
국내는 일부 연구소 및 IT기업을 중심으로 기계학습 기술개발이 진행 중이며, 정부 연구개발(R&D) 과제를 확대해 나갈 예정이다.

삼성전자는 애플의 시리와 유사하게, 음성인식 기술을 적용한 'S보이스'에 이어 '빅스비'를 선보이며 자연어처리기술 기반 음성인식 기술을 개발하는 데 초점을 맞추고 있다. 매주 국가별로 사투리, 축약어를 포함한 수천 개 문장을 테스트하며 학습 데이터를 확보하고 있으며, 노이즈가 발생하는 환경에서도 음성인식 정확도를 유지할 수 있는 기술을 개발 중에 있다. 또한, 인공지능 기술을 미래의 핵심 서비스로 간주하여, 미국 실리콘밸리 소재 인공지능(AI) 플랫폼 개발 기업인 비브 랩스사(VIV Labs Inc.)를 인수하고 자사 스마트폰 플래그십 모델에 전용 버튼을 할당하는 등 서비스 경험 향상을 위해서 전력을 기울이고 있다. 삼성전자는 딥러닝 응용 프로그램 개발을 위한 분산형 플랫폼 '베레스(Veles)'를 오픈소스로 공개하였다. 개방형 범용 병렬 컴퓨팅 프레임워크인 '오픈(Open)CL' 또는 GPU에서 수행하는 알고리즘을 C 프로그래밍 언어를 비롯한 산업 표준 언어를 사용해 작성할 수 있도록 하는 GPGPU 기술인 '쿠다(CUDA)' 등을 사용하며, 파이썬 등을 통해 플로 기반의 프로그래밍이 가능하여 개발자들의 활용률을 높이고자 하였다.

네이버는 기계학습 기술을 이용하여 엔드라이브(Ndrive)나 검색어 자동 완성 기능 등을 제공하고 있다. 엔드라이브에 적용된 기계학습 기술은 구글 포토와 같이 사용자가 사진을 업로드하면 동물, 음식, 텍스트 등으로 카테고리를 나누어 자동으로 분류를 해준다. 검색어 자동 완성 기능 역시 기계학습 기술에 의해 만들어졌으며, 사용자가 검색창에 첫 글자만 입력해도 과거 데이터를 분석해 사용자가 원할 만한 내용을 예측해서 알

려주고 있다. 네이버 번역 서비스인 파파고 또한 인공신경망 번역 기술을 적용하여 정확도를 많이 높이고 있으며, 현재 한국어에 대해서는 구글 번역 서비스에 필적할 만한 수준을 보여 주고 있다.

[그림 2-17] 통역앱 파파고



자료: 구글 스토어. (2017). http://store.google.com 파파고(papago) 상세 설명 이미지에서 2017. 11. 29. 인출.

2017년 3월 독자적인 인공지능 기술 '클로바(Clova)'를 처음 공개한 것을 시작으로 최근 두 달간 인공지능 스피커 '웨이브', 차량용 인공지능 기기 'IVI', 인공지능 로봇 '페이스(FACE)'와 같은 분야별 프로젝트를 연이어 발표하고 있다(성호철, 2017). 웨이브는 미국 아마존의 AI 스피커 '에코'처럼 이용자의 음성을 듣고 음악을 틀어 주거나 날씨·교통 등 각종 정보를 알려주는 기기다. 2017년 8월에 선보인 IVI는 차량 계기판 상단에 설치하는 태블릿PC 모양의 기기다. 운전자가 말하면 내비게이션·날씨·일정·음악 등을 실행해 보여 준다. 스마트폰 형태의 IVI '어웨이'는 차량 공유 업체 그린카를 통해 시범운영하고 있다.

[그림 2-18] 네이버의 인공지능 스피커: 웨이브

자료: WAVE. (2017). http://store.clova.ai/ko/product/wave 웨이브 상세 설명 이미지에서 2017. 11. 29. 인출.

네이버는 기계학습을 포함한 첨단 기술의 빠른 확보를 위해 작년에만 15개 벤처 기업에 1000억 원이 넘는 지분 투자를 한 상황이다.

카카오 또한 창업자가 직접 인공지능 자회사 카카오브레인의 대표를 맡을 정도로 기계학습 기술 확보에 사력을 다하고 있다. 메신저서비스인 카카오톡과 인터넷포털 다음, 디지털 음원 자회사 멜론 등을 통해 확보된 국내 이용자들에 대한 방대한 데이터를 바탕으로 인공지능 소프트웨어를 개발하고 이를 활용해 인공지능 스피커 등 각종 신규 서비스를 선보일 계획이다.

현대자동차는 2010년부터 자율주행 자동차 기술 확보를 위한 노력을 해오고 있으며, 2012년에 고속도로 주행 지원 시스템 개발을 완료하고, 2017 CES에서 대도심 야간 자율주행에 성공하였다. 이미지, 음성인식기술과 딥러닝에 기반을 둔 고속도로 주행 지원 시스템은 차선이탈 경보,

차선 유지 지원, 후측방 경보, 차량 속도 유지, 자동 긴급제동 등의 기능이 포함된다. 또한 운전자가 운전 중에도 자동차를 가장 편하게 조작할 수 있도록 음성인식 기술 확보에도 많은 노력을 기울이고 있으며, 2018년에 인공지능(AI)의 딥러닝 음성인식 장치를 탑재한 차량을 출시할 예정이다.

클디(cldi)는 딥러닝 기술을 이용해 이미지로부터 관련 정보를 인식하는 기술을 개발하는 스타트업이다. 클디는 2014년 이미지 인식대회인 ILSVRC(Imagenet Large Scale Visual Recognition Challenge)에 참가해 물체분류 및 위치인식 부문 세계 7위를 차지했으며, 2015년 주목해야 할 한국의 스타트업 22개 기업 중 하나로 해외에서 소개되었다.

유빅은 국제소송 전자증거개시제도 서비스를 제공하는 기업으로, 기계학습 기반 예측코딩 기술 도입으로 이메일을 통한 기술유출을 차단할 수 있는 솔루션을 개발하였다. 예측코딩 기술을 활용해 이메일에 담긴 기술정보 유출, 리베이트, 카르텔, 부정회계, 횡령 등 기업 부정 사건의 위험 징후를 감지할 수 있으며, 이메일을 계정별로 분석하고, 분석을 마친 이메일을 위험 징후의 세 단계로 분류하여 사용자가 위험 징후를 더욱 정확히 예측할 수 있도록 한다.

4. 분야로 살펴본 기계학습 기법의 활용 사례[5)]

앞서 기술한 국내·외 기계학습 기법 활용 사례는 각 기업과 기술 중심으로 살펴보았다. 교통, 농업, 보건, 교육, 문화예술, 환경 분야로 나누어 우리 삶에 미치는 기계학습 영향은 어떠한지 살펴보자.

교통 분야는 이제 앞서 기술한 것처럼 자율주행까지 가능하고 어떤 윤리적 문제가 발생할지에 대한 법적 논의까지 하는 단계이다. 기계학습의

5) 고학수(2017), pp. 77-78을 포괄적으로 인용하여 재구성함.

진화는 운송과 화물 선적 분야에서 획기적인 변화를 줄 뿐만 아니라 보다 안전하고 저렴하면서도 친환경적인 교통수단을 제공해 줄 수 있을 것이다.

농업 분야에서는 기계학습 기법을 활용하여 반복적인 프로세스의 자동화 작업 부분에서 생산성과 노동의 효율성을 지속적으로 증가시킬 수 있다. 또한, 토양조건 및 과거의 기상 패턴 등의 다양한 요인들을 분석하여 기후변화나 자연재해의 영향을 예측할 수 있다. 기계학습의 역량을 보여주는 사례로, 호주의 농부들은 '농장 봇(farmbot)'을 훈련시켜 가축을 기르고 건강을 모니터링 한다. 그리고 일본의 농부들은 기계학습과 구글의 텐서플로(Tensorflow)를 활용하여 오이를 크기, 모양, 기타 속성별로 자동 분류하는 방식에 적용하여 생산성을 높이고 있다.

보건 분야는 다른 분야보다 기계학습으로 인한 우리 삶의 변화가 가장 클 것이다. 예를 들어, 당뇨병성 망막증(diabetic retinopathy)은 당뇨병의 합병증이고 전 세계적으로 가장 빠르게 증가하고 있는 실명의 원인이다. 이 병은 초기에 발견되는 경우 큰 문제가 되지 않지만 개발도상국에서는 당뇨병성 망막증의 선별검사와 진단을 할 수 있는 안과의사 수가 충분하지 않은 실정이다. 이런 상황에서 기계학습 기법을 활용하면 더 많은 사람들이 당뇨병성 망막증을 진단받을 수 있을 것이다(Gupta, Karandikar, 2015). 앞서 기술한 딥마인드(DeepMind)와 베릴리(Verily)를 포함한 다른 회사들도 기계학습을 이용하여 질병의 진단, 예방 치료를 개선하기 위한 다양한 연구를 진행하고 있다. 최근에 미국의 방사선 전문의들은 텐서플로를 사용하여 검사 결과에서 파킨슨병의 징후를 찾아내는 방법에 관한 논문을 발표하였다. 기계학습은 정확한 진단을 내리는 데 도움을 줄 뿐만 아니라, 맞춤화되고 자동화된 서비스를 제공하여 보다 효율적으로 치료계획을 수립하고, 투약을 관리하며 노인 환자들을 간호하는 데 큰 기여를 할 수 있을 것이다.

교육 분야에서의 기계학습은 스마트 어시스턴스(smart assistance)와 개별 학습자에게 맞춤화된 콘텐츠를 제공하여 학습을 도울 수 있다. 이미 온라인 공개 수업(Massive Open Online Course)에서는 비디오 강좌, 토론 게시판, 자동 채점 시스템을 활용하고 있으며 기존의 대학 수업 및 원거리 학습자들에게 큰 도움을 줄 수 있다. 그리고 학습 분석(learning analytics)분야와 교과과정 자료 선정 과정에서 중요한 역할을 할 수 있을 것이다.

문화예술 분야도 기계학습을 이용한 실험적인 시도가 이루어지고 있다. 구글 마젠타(Google Magenta)연구 프로젝트는 예술가들과 창작자들이 기계학습 기법을 활용하여 음악, 비디오, 이미지 및 텍스트 생성에서 최첨단 기술을 개발하는 데 도움을 주고 있다. 구글 마젠타는 궁극적으로 기술에 관심이 있는 예술가, 코딩 전문가, 최첨단 컴퓨터 기술 및 인공 신경망을 이용하여 실험적 시도를 하는 기계학습 전문가들이 모일 수 있는 공동체를 만들고자 하는 목표를 가지고 있다.

환경 분야에서의 기계학습은 전기시스템의 에너지 효율성을 증가시켜 에너지 소비 문제와 기후변화 문제의 해결에 도움을 줄 수 있다. 지난 10년간 구글의 주된 목표 중 하나는 에너지 사용량을 줄이는 것이었는데, 실제로 구글은 효율성 높은 서버들을 사용하고, 새로운 데이터 센터 냉각 기술을 활용하며, 신재생 에너지만을 사용함으로써 에너지 사용량을 상당히 감소시켰다. 최근에는 딥마인드의 기계학습을 구글 데이터 센터에 적용하여 에너지 사용량을 40퍼센트가량 줄였다. 이는 구글의 데이터 센터뿐만 아니라 구글의 클라우드를 사용하는 다른 회사들도 이런 기술을 통해 에너지 효율성을 제고하고 탄소 배출량을 줄일 수 있을 것이다.

이렇듯, 기계학습은 다양한 분야에서 직접적으로 우리 삶에 영향을 미치며 빠르게 발전해 나갈 것이다.

제3절 딥러닝(Deep Learning) 기법 동향 분석

최근 들어 빅데이터의 출현, 학습 기법의 개선, 컴퓨팅 파워의 증진으로 인해서 기존의 기계학습 및 ANN(Artificial Neural Network)이 가지고 있던 기존의 한계를 풀어나가며 다양한 문제들에 대해 획기적인 성능 향상을 이뤄냈고, 이에 따라 기계학습 및 딥러닝에 대한 관심이 더욱 커지고 있다. 이처럼 강력한 기계학습 및 딥러닝이 앞으로 다가올 4차 산업혁명 시대를 견인할 구동요인으로 조명되고 있다. 이번 절에선 기존의 기계학습 및 딥러닝 기법들에 대한 소개뿐만 아니라 최근 기계학습 및 딥러닝 기법들의 개념 및 기술적 동향들에 대해서 소개하고자 한다.

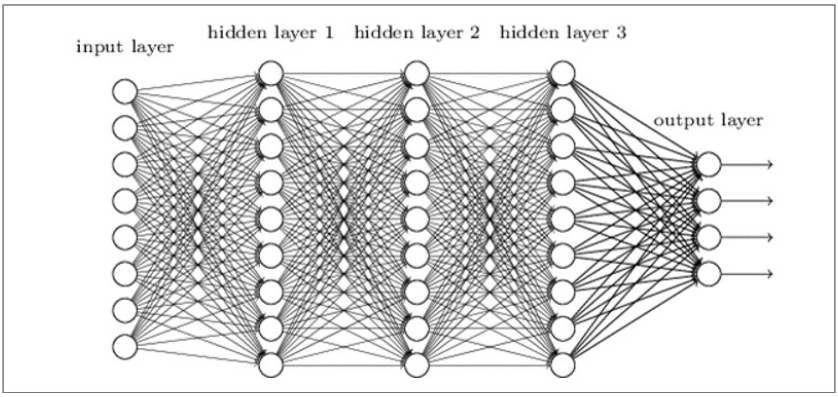
최근 여러 학문 분야 및 산업 분야에서 적용되는 딥러닝 기반의 기계학습들의 개념, 구조 그리고 기술들이 활용는 최신 동향들에 대해서 소개하고자 한다. 딥러닝 기반의 기계학습 알고리즘들은 이미지처리는 물론, 자연어처리, 음성인식, 로봇 분야 등에서 매우 다양하게 활발히 연구 중이다. 딥러닝 연구의 폭발적인 성장에는 이러한 '연구 분야의 통합' 이 큰 기여를 하고 있다. 이전에는 각각의 도메인에서 따로 연구가 진행되었지만 '딥러닝' 이라는 한 가지 주제가 이슈화되다 보니 연구의 다양성이 증가되고 연구의 양과 속도 또한 빠르게 발전을 이루고 있다.

1. 딥러닝 기반의 기계학습 기법: DNN(Deep Neural Network)

딥러닝이란 기계학습의 한 분야이다. 기존의 기계학습과 차이점이 있다면, 기계학습에서 기계가 학습하기 위해 주어진 데이터에서 특징을 추출하는 과정에 여전히 사람이 개입하지만 딥러닝은 주어진 데이터를 그대로 입력데이터로 활용한다는 점이다. 사람이 생각한 특징을 학습하는 것이 아니라 데이터 자체에서 중요한 특징을 기계 스스로 학습하는 것이다.

그래서 보통 딥러닝을 end-to-end machine learning 이라고 한다. 딥러닝의 경우 신경망의 모델이 입력층, 은닉층, 출력층으로 구성된다. 기존의 기계학습들은 은닉층(hidden layer)의 개수가 적어서 'Shallow learning' 이라고 하고, 은닉층의 개수가 많은 신경망 모델을 가리켜 'Deep learning'이라고 한다. 딥러닝의 가장 큰 장점 중 하나로는 기존의 shallow learning으로는 풀 수 없는 복잡한 문제들을 풀 수 있다는 것이다. [그림 2-19]는 DNN의 기본 구조에 대해서 나타내는 그림이다. 아래 그림처럼 여러 개의 은닉층으로 구성되어 있다.

[그림 2-19] DNN의 기본구조



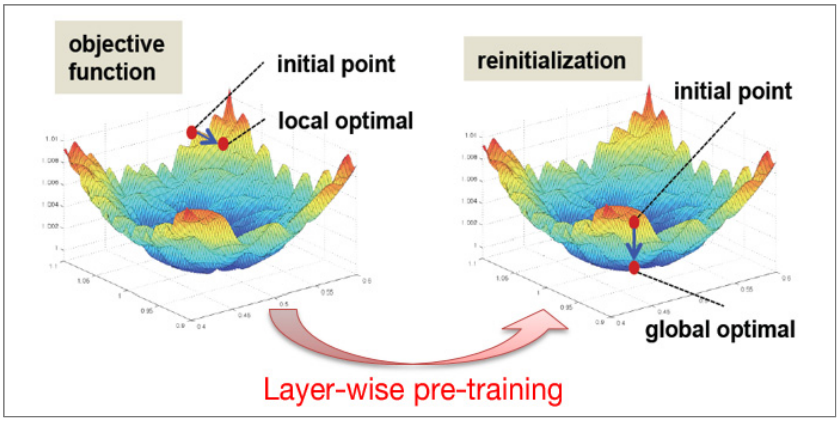
자료: Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science, 2006, Jul 28: 313.

딥러닝의 기술이 발전하는 데 중요한 기법들이 몇 가지 있는데 그 중 하나가 'pre-training' 기법이다. 2006년 힌턴 교수의 사이언스 논문 "Reducing the Dimensionality of Data with Neural Networks"에서 제안된 방법이다.

딥러닝 학습의 주요 문제점 중 하나였던 gradient vanishing을 파라미터의 초기값를 효과적으로 설정해줌으로써 해결하는 방법이다. DNN

은 RBM(Restricted Boltmann Machines)을 stacking한 구조의 DBN (Deep Belief Network)과 같이 layer-wise하게 비지도학습으로 가중치(weight)의 초기값을 잘 설정하여 해결하는 방법 또한 존재한다.

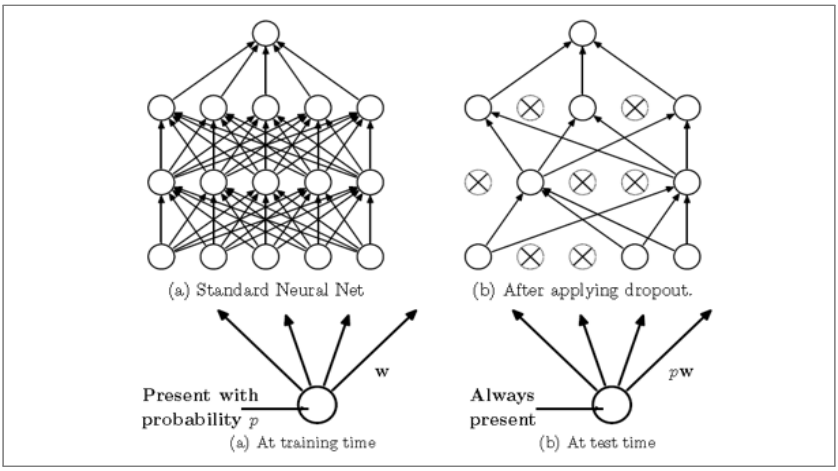
[그림 2-20] Pre-training의 효과



자료: Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science, 2006, Jul 28: 313.

DNN에서 문제를 해결해 나갈 때에 최적화 과정이 필요하다. 가장 기본적인 최적화 기법으로는 Gradient descent 기법이 있다. learning rate와 출력층 및 은닉층의 gradient 값을 이용하여 파라미터 값들을 업데이트 하는 방법이다.

〔그림 2-21〕 Drop-out의 기본 구조에 대한 개념

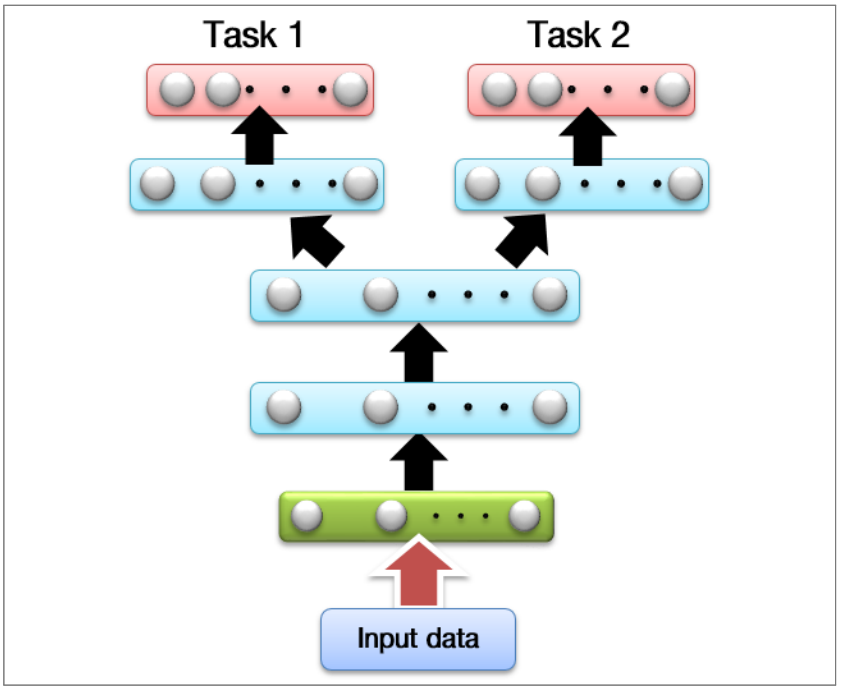


자료: Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science, 2006, Jul 28: 313.

또한 DNN에서 overfitting 방지를 위한 정규화 기법을 사용한다. Overfitting 방지를 위한 대표적인 정규화 기법은 'Drop-out' 이라는 기법이다. 〔그림 2-21〕은 Drop-out 기법의 기본 개념에 대한 구조이다. Drop-out의 경우 학습 시 random하게 일정한 비율(p)의 node들을 비활성화함으로써 다양한 조합의 DNN을 학습한다.

다양한 조합의 DNN들이 voting 효과를 가져와 하나의 DNN에 과도하게 학습되지 못하도록 강력한 정규화 기법처럼 작용한다. 테스트 시에는 node들을 비활성화 하지 않고 파라미터 값에 p를 곱해서 사용한다.

[그림 2-22] Multi-task learning의 기본 구조



자료: Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science, 2006, Jul 28: 313.

DNN의 정규화 기법 중 또 다른 기법으로는 'Multi-task learning' 이라는 기법이 있다. Multi-task learning 이란 이름에서 알 수 있듯이 추정하고자 하는 task(task1)외에 다른 task(task2)를 추가적으로 학습시키는 방법이다. 추정하고자 하는 task와 어느 정도 상관도(correlation)가 있는 task를 함께 학습할 경우 추가된 task가 원래 추정하고자 했던 task의 학습에 영향을 끼쳐 추정 성능이 향상된다. 이러한 현상을 정규화 관점에서 보면 task 2를 추가함으로써, 학습이 task 1에 과도하게 진행되지 않도록 정규화해주는 역할을 한다고 볼 수 있다.

DNN을 학습 시키는 툴(tool)로는 여러 가지가 사용된다.

- ○ Theano: 몬트리올대학(Yoshua Bengio 교수팀)이 개발한 툴로써, Bengio 교수를 중심으로 관련 커뮤니티가 매우 활성화 되어 있으며 Python 기반의 툴임.
- ○ Caffe: 캘리포니아대학 버클리(Berkely Vision and Learning Center)가 개발하였고 컴퓨터 비전 분야에 특화되어 있으며, C/C++ 및 Python 기반의 툴임.
- ○ Torch: 뉴욕대학(Yann LeCun 교수팀, 페이스북에서 주로 사용)이 개발한 툴이며, Python 기반의 툴보다 속도가 다소 빠른 것으로 알려져 있고, C/C++ 기반의 툴임.
- ○ Tensorflow: 구글이 개발하였으며, 다른 툴들에 비해 다소 늦게 공개되었으나 쉬운 사용법과 빠른 업데이트 등으로 영향력이 점점 커지고 있는 툴이며, C/C++ 및 Python 기반의 툴임.

현재 DNN은 다양한 분야에서 좋은 성능을 내며 사용되고 있다. 영상, 이미지 분야 및 음성분야, 바이오분야 등 여러 분야에서 기존의 패러다임을 바꿀 정도의 고무적인 성능을 보이고 있다. 이에 따라 기존의 DNN 뿐만 아니라 새로운 구조들이 등장하게 되며 그 새로운 구조들에 대한 실험 결과 및 성능들이 빠른 흐름으로 진행되고 있다. 최근 딥러닝의 다양한 구조 및 동향들에 대해서 더욱 자세히 알아보고자 한다.

2. 딥러닝 기반의 기계학습 기법: CNN(Convolutional Neural Network)[6)]

CNN은 하나 또는 여러 개의 컨볼루션 레이어와 그 위에 올려진 일반적인 인공 신경망 레이어들로 이루어져 있다. CNN의 이러한 구조는 2차원 입력 데이터에 적합한 구조이며, 따라서 음성이나 영상 분야에서 좋은 성능을 보여준다. CNN의 구조를 자세히 보면 기존의 신경망과 비슷하나 컨볼루션을 통해 유의미한 특징을 추출한다는 점에서 다르다. 또한, CNN은 다른 딥러닝들의 구조처럼 여러 레이어들이 스태킹(stacking)되어 있으며, 레이어들의 종류는 다양하나, 기본적인 레이어들을 보면 다음과 같이 세 가지 종류라고 볼 수 있다.

○ 컨볼루션 레이어(convolution layer): 컨볼루션 연산을 통해 특징들을 추출하는 레이어

○ 풀링 레이어(pooling layer): 입력 공간을 추상화하는 레이어, 예를 들어 영상 데이터의 경우 픽셀의 수가 많으면 서브샘플링(sub-sampling) 등의 과정을 통해 차원 축소(dimensionality reduction)의 효과를 얻음.

○ 전방향 레이어(feedforward layer): 최상위 레이어들에서 마지막에 적용되며, 하위 레이어에서 전달된 특징들을 분류하는 역할을 함.

기본적으로 학습 모델은 추론과정과 학습과정을 가지게 된다. 추론과정은 학습된 내용을 기반으로 새로운 입력에 대해 답을 얻어내는 과정이며, 학습은 주어진 학습 데이터를 기반으로 최적의 추론을 수행하기 위해 추론 구조 또는 학습 파라미터(가중치)들을 설정하는데, 간단히 말하면

6) 강대기(2016). pp. 15-16을 포괄적으로 인용하여 재구성함.

귀납적으로 배우는 단계로 볼 수 있다.

CNN의 전방향 레이어의 경우, 일반 전방향 다층 퍼셉트론(Feedforward Multilayer Perceptron: MLP)처럼 추론을 위해서는 하위 레이어의 출력에 대해 가중치를 곱한 것에 대한 바이어스의 합을 시그모이드(sigmoid) 함수와 같은 활성 함수에 통과시킨 값이 된다. 학습의 경우도 전방향 MLP처럼 실제 값과 예측 값에서 구해지는 에러 값에 대해 가중치, 바이어스 등을 그래디언트(gradient)를 구해 업데이트하는 방향으로 이루어진다.

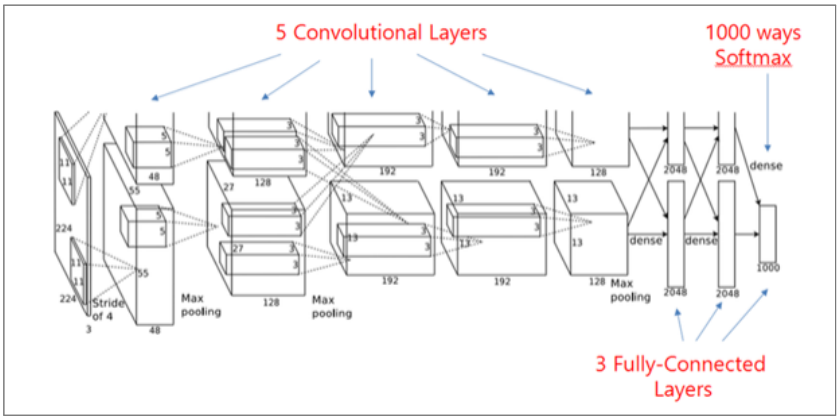
CNN의 컨볼루션 레이어의 경우, 전방향 레이어와 크게 다르지 않으나, 하나의 필터를 하나의 이미지에 적용하는 역할을 하므로 하나의 레이어에 대해 각 노드들에 특정 가중치는 모두 동일하게 된다. 하나의 이미지가 여러 픽셀들에 대해 동일한 필터를 적용하는 것이 일반적인 전처리 과정으로 각각의 픽셀에 대해 동일한 가중치 집합이 적용된다고 보면 될 것이다. 또 한 가지 다른 점은 하위 레이어 노드에서 상위 레이어 노드의 연결 형태를 보면, 완전히 연결된 그래프(fully connected graph)의 형태가 아니다. 즉, 필터가 이미지 전체에 적용되는 것이 아니라, 이미지의 일부에 적용되는 형태인 것이다.

풀링 레이어는 서브샘플링을 수행하는 방법으로 여러 개의 픽셀들을 하나의 픽셀로 사상(mapping)시키는 추상화를 수행한다. 이론적으로는 여러 가지 방법의 서브샘플링이 수행될 수 있으나 최대값을 반환하는 형태의 맥스 풀링 기법이 주로 사용된다.

지금까지는 CNN의 기본적인 개념과 구조에 대해서 살펴보았다. CNN은 데이터로부터 자동으로 피처(feature)를 학습하는 대표적인 알고리즘이라고 할 수 있다. 기존의 기계학습을 통해 데이터를 학습하기 위해선 먼저 날 것의 데이터를 피처로 가공하는 과정이 필요하다. 딥러닝,

특히 CNN은 이러한 피처를 데이터로부터 매우 효율적으로 학습한다. 현재 CNN은 2012년 ILSVRC 이미지인식 대회에서 힌턴 교수 팀의 AlexNet이 놀라운 성능 개량을 보임으로써 CNN의 폭발적인 연구 성장이 이어져 왔다. 지금부터는 CNN에 대한 이해를 좀 더 깊게 하기 위해서 그동안 ILSVRC 라는 이미지인식 대회에서 우수한 성능을 보인 CNN 구조들에 대해서 살펴보고자 한다.

[그림 2-23] AlexNet의 기본 구조



자료: Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS'12 Proceedings of the 25th International conference on Neural Information Processing Systems, Vol(1), 1097-1105.

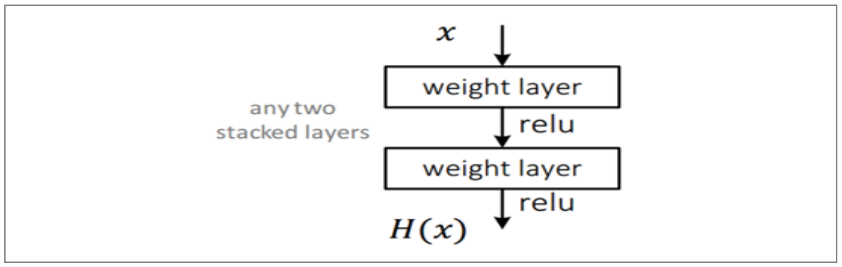
2012년 ILSVRC(ImageNet Large Scale Visual Recognition Challenge) 대회에서의 우승은 캐나다 토론토 대학이 차지했으며, 당시 논문의 첫 번째 저자가 Alex Khrizevsky 이다. 그의 이름을 따, 그들이 개발한 CNN 구조를 AlexNet라고 부른다. AlexNet은 총 5개의 convolution layers와 3개의 full-connected layers로 구성이 되어 있으며, 맨 마지막 full-connected layer는 1000개의 카테고리로 분류를 위

해 활성함수로 softmax 함수를 사용하고 있다.

AlexNet에서는 pooling 기법으로 max pooling 기법을 쓴다. max pooling은 윈도에서 최대값을 갖는 픽셀을 선택하는 기법이다. 또한 활성함수로 sigmoid가 아닌 ReLU를 사용하여 학습속도를 줄일 뿐만 아니라 back propagation이 간단해지는 이점을 가지고 있다. Overfitting에 대한 문제를 해결하기 위한 대표적인 방법 중 하나가 데이터의 양을 늘리는 것이다. 그러나 학습 데이터를 늘리는 것이 쉽지 않으며, 학습 데이터가 늘어나면 학습 시간이 길어지기 때문에 효율성을 반드시 고려해야 한다. 따라서 data argumentation 이라는 기법을 사용해서 over-fitting의 문제를 극복하였다.

ResNet(Residual Net)은 2015년 ILSVRC에서 우승을 한 알고리즘이며 설계자는 'Kaiming He' 라는 사람이다. ResNet은 Activation을 직접적으로 변환하는 weight를 학습하는 게 아니라 입력(input)과 출력(output)의 차이인 잔차(residual)를 학습시키는 방식이다. 즉 기존 AlexNet 같은 경우 Convolution을 통해 Feature maps를 Transformation 시켜 왔지만 ResNet 같은 경우 입력(input)에 Additive한 값을 Convolution으로 학습시킨다. 구체적으로 ResNet 설계 팀은 망을 100 layer 이상으로 깊게 하면서, 깊이에 따른 학습 효과를 얻을 수 있는 방법을 고민하였으며, 그 방법으로 Residual Learning 이라는 방법을 발표하였다(인하대, 2017). Residual 이라는 단어를 사용한 이유는 아래 구조를 살펴보면 파악할 수가 있다. 먼저 아래 구조와 같은 평범한 CNN 망을 살펴보도록 한다. 이 평범한 망은 입력 x를 받아 2개의 weighted layer를 거쳐 출력 $H(x)$를 내며, 학습을 통해 최적의 $H(x)$를 얻는 것이 목표이며, weight layer의 파라미터 값이 결정된다.

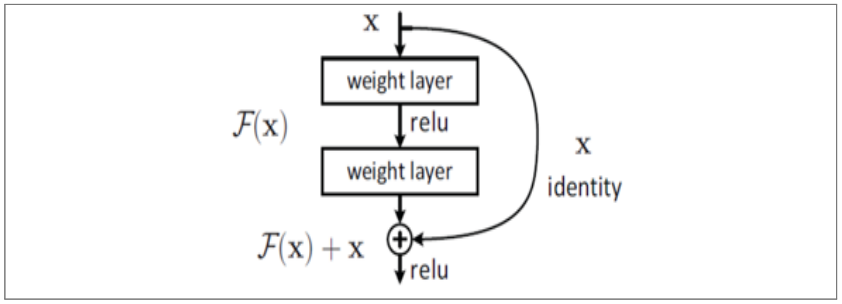
〔그림 2-24〕 기본적인 구조의 CNN 망



자료: He, K. (2015). Deep Residual Learning. Conference on Computer Vision and Pattern Recognition. Boston.

그런데 〔그림 2-24〕의 기본적인 CNN망의 구조에서 $H(x)$를 얻는 것이 목표가 아니라 $H(x)-x$를 얻는 것으로 목표를 수정한다면 즉, 출력과 입력의 차를 얻을 수 있도록 학습을 하게 된다면, 2개의 weighted layer는 $H(x)-x$를 얻도록 학습이 되어야 한다. 여기서 $F(x)=H(x)-x$ 라면 결과적으로 출력 $H(x)=F(x)+x$ 가 된다. 그러므로 〔그림 2-24〕 블록도는 〔그림 2-25〕처럼 바뀌며, 이것이 바로 Residual Learning의 기본 블록도가 된다.

〔그림 2-25〕 ResNet의 기본구조



자료: He, K. (2015). Deep Residual Learning. Conference on Computer Vision and Pattern Recognition. Boston.

달라진 점이 있다면, 입력에서 바로 출력으로 연결되는 shortcut 연결이 생기게 되었으며, 이 shortcut은 파라미터가 없이 바로 연결이 되는 구조이기 때문에 연산량 관점에서는 덧셈이 추가되는 것 외에는 차이가 없다.

이렇게 기존 구조에서 관점이 바뀜에 따라 꽤 많은 효과를 얻을 수 있다. 예전에는 $H(x)$를 얻기 위한 학습을 했다면, 이제는 $H(x)-x$를 얻기 위한 학습을 하게 되며, 최적의 경우라면 $F(x)$는 0이 되어야 하기 때문에 학습할 방향이 미리 결정이 되어, 이것이 pre-conditioning 구실을 하게 된다. $F(x)$가 거의 0이 되는 방향으로 학습을 하게 되면 입력의 작은 움직임(fluctuation)을 쉽게 검출할 수 있게 된다. 그런 의미에서 $F(x)$가 작은 움직임, 즉 나머지(residual)를 학습한다는 관점에서 residual learning 이라고 불리게 된다. 또한 입력과 같은 x가 그대로 출력에 연결이 되기 때문에 파라미터의 수에 영향이 없으며, 덧셈이 늘어나는 것을 제외하면 shortcut 연산량 증가는 없다. 그리고 몇 개의 레이어를 건너뛰면서 입력과 출력이 연결이 되기 때문에 forward나 backward path가 단순해지는 효과를 얻을 수 있다.

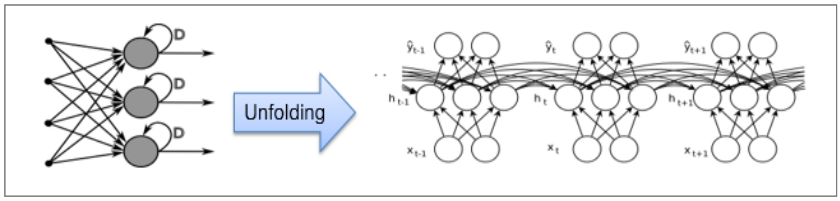
3. 딥러닝 기반의 기계학습 기법: RNN(Recurrent Neural Networks) & LSTM(Long Short Term Memory) networks

RNN은 사람이 어떻게 생각하는지를 초점에 두고 신경망이 설계되었다. 사람은 생각을 매순간마다 처음부터 다시 시작하지 않는다. 기존의 기억하던 것을 버리고 처음부터 생각을 다시 시작하지 않는다. 사람의 생각에는 지속성이 있다. 그러나 전통적인 신경망들은 이런 일을 할 수 없었다. 그것이 전통적 신경망들의 주요 단점이었다. DNN이나 CNN은 이

전 정보와의 correlation을 충분히 고려해 주지 못하는 단점을 가지고 있었다. RNN은 내부에 루프를 가진 네트워크이다. 즉 그 루프는 정보가 지속되는 것을 도와주는 역할을 한다. RNN은 다음 정보가 신경망에 통과될 때 이전 정보 값이 다시 recurrent하게 고려되어 연속적인 성질을 가진 데이터에 대해 학습할 때 효과적이다. 그러나 정보 간의 연관성이 고려되지 못하고 시간이 지날수록 이전 정보의 값이 사라져 버리는 vanishing gradient 문제가 있다.

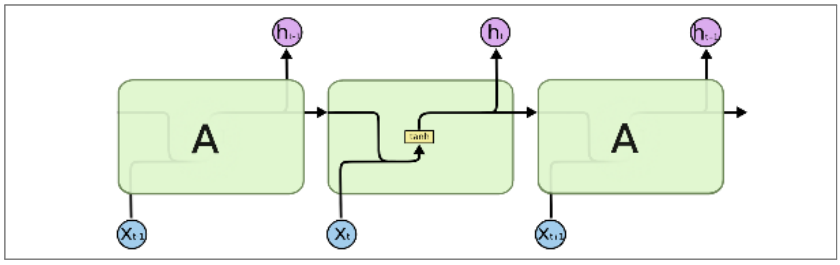
LSTM은 장기 의존성을 학습할 수 있는 특별한 종류의 순환 신경망이다.

〔그림 2-26〕 RNN의 recurrent 성질/ RNN을 시간에 따라 펼친 그림



자료: Colah. (2015). Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs에서 2017. 11. 29. 인출.

〔그림 2-27〕 RNN에서 반복되는 모듈들의 구조



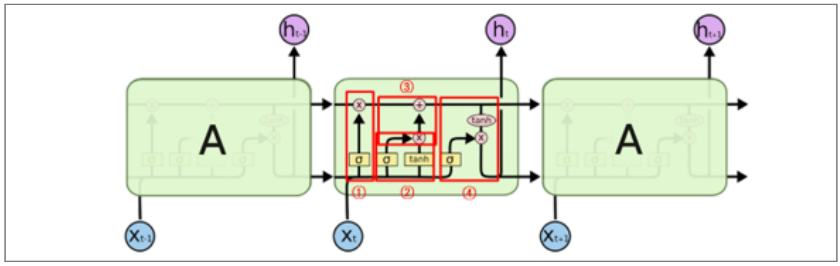
자료: Colah. (2015). Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs에서 2017. 11. 29. 인출.

연속적인 데이터를 학습할 때 강력한 성능을 나타내는 RNN에서의 gradient vanishing 문제를 해결하고자 설계되었다. 기존의 RNN은 사

슬 형태의 반복되는 신경망 모듈들을 가지고 있다. 이 반복되는 모듈은 한 개의 tanh 층 같은 매우 간단한 구조를 가지고 있다.

LSTM 또한 사슬 같은 구조를 가진다. 그러나 반복되는 모듈은 다른 구조를 지닌다. RNN에서의 gradient vanishing 문제를 gradient 정보를 저장할 수 있는 셀(cell) 개념과 3가지 gates(forget gate, input gate, output gate) 개념을 도입하여 해결하였다.

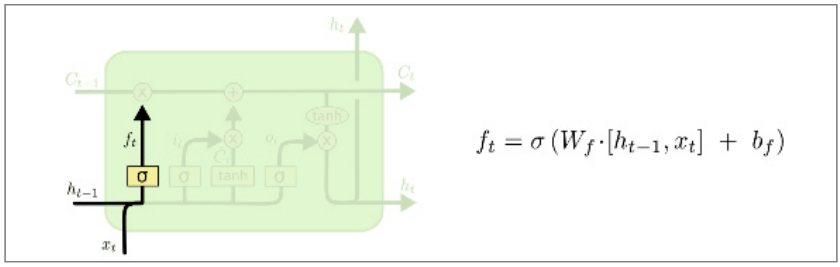
[그림 2-28] LSTM의 기본 구조



자료: Colah. (2015). Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs에서 2017. 11. 29. 인출.

LSTM에서 첫 단계는 셀 상태에서 어떤 정보를 버릴지 결정하는 것이다. 이를 'Forget gate'라고 하며, 시그모이드 층에 의해서 결정된다.

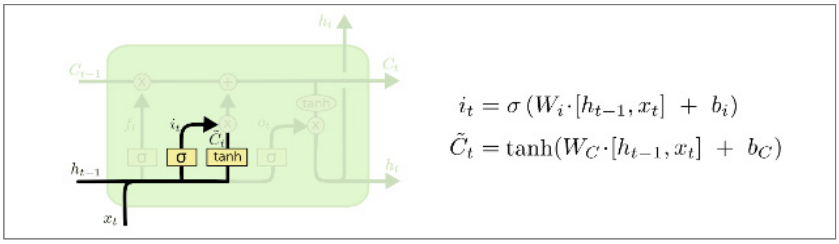
[그림 2-29] Forget gate의 구조 및 수식



자료: Colah. (2015). Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs 에서 2017. 11. 29. 인출.

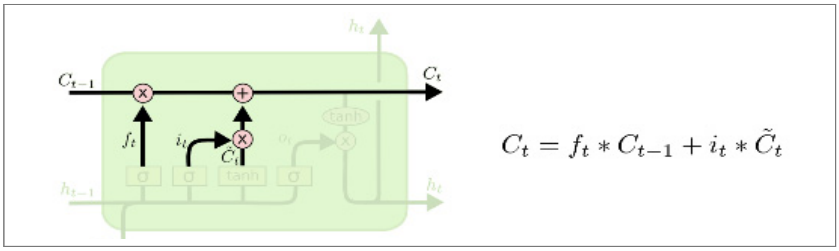
다음 단계에서는 어떤 새로운 정보를 셀 상태에 저장할지 결정한다. 이 단계는 두 부분으로 구성된다. 첫째, “입력(input) 게이트 층” 이라 불리는 시그모이드 층은 어떤 값들을 갱신할지 결정한다. 다음 tanh 층은 셀 상태에 더해질 수 있는 새로운 후보 값들의 벡터 C_t를 만든다. 그다음 단계에서, 셀 상태를 갱신할 값을 만들기 위해 이 둘을 합치게 된다.

[그림 2-30] Input gate 의 구조 및 수식



자료: Colah. (2015). Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs에서 2017. 11. 29. 인출.

[그림 2-31] Cell 의 구조 및 수식

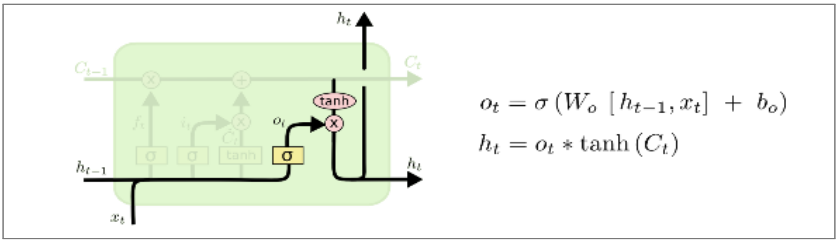


자료: Colah. (2015). Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs에서 2017. 11. 29. 인출.

이번 단계에서는 낡은 셀 상태 C_{t-1}을 새 셀 상태 C_t로 갱신하는 단계이다. 이전 상태 C_{t-1}에 f_t를 곱한다. f_t는 우리가 전에 계산한 forget gate의 출력이다. 그런 다음 $i_t * \tilde{C}_t$를 더한다. 이것이 각 상태 값을 우리가 얼만큼 갱신할지 결정한 값으로 크기 변경한(scaled) 새 후보 값들이다.

마지막 단계로 무엇을 출력할지를 결정한다. 이 출력은 셀 상태에 기반을 두지만 여과된 버전이다. 우선, 시그모이드 층을 동작시킨다. 그 시그모이드 층은 셀 상태에서 어떤 부분들을 출력할지 결정한다. 그런 다음 값이 -1과 1 사이 값을 갖도록 셀 상태를 tanh에 넣는다. 그리고 오직 결정한 부분만 출력하도록, tanh 출력을 다시 시그모이드 게이트 출력과 곱한다.

[그림 2-32] Output gate의 구조 및 수식



자료: Colah. (2015). Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs에서 2017. 11. 29. 인출.

LSTM을 통해서 기존 RNN의 문제들을 해결할 수 있게 되었고 성능적 측면에서도 보다 좋은 결과를 성취할 수 있게 되었다.

4. 딥러닝 기반의 기계학습 기법: RL(Reinforcement learning)[7)]

강화학습은 확률적 의사결정 문제를 푸는 방법론들을 지칭한다고 볼 수 있다. 강화학습은 지도학습에 속한 모델로 분류하기도 하고 강화학습 자체의 독립적인 영역으로 분류하기도 한다. 지도학습으로 분류되는 이유로는 학습 중에 사람을 포함한 환경으로부터 피드백, 즉 지도를 받기 때문이다. 한편 독립적으로 분류되는 이유는 강화학습이 가지고 있는 최

7) 최성준, 이경재(2017)를 포괄적으로 인용하여 재구성함.

적의 의사결정 프로세스가 지도 학습의 대표적인 방식인 레이블 기반을 통한 판별식을 구하는 방식과는 구별되는 학습 모델이기 때문이다. 강화학습은 시행착오 과정을 거쳐 학습하기 때문에 사람의 학습방식과 매우 유사하다. 이러한 이유로 구글 딥마인드가 설계한 알파고의 핵심 알고리즘도 강화학습을 기반으로 한다. 따라서 강화학습은 인공지능 분야에 가장 가까운 모델이라 할 수 있다. 지도학습과는 다르게 강화학습은 학습 데이터가 주어지지 않는다. 그 대신 강화 학습 문제에 주어지는 것은 보상함수이다. 강화학습을 푼다는 것의 정의는 미래에 얻어질 보상 값들의 평균을 최대로 하는 정책 함수를 찾는 것이다. 또한 강화학습을 풀기 위해서 연구자들은 수학적 모델을 하나 차용했는데, 그것이 바로 마코프 의사결정 과정(Markov decision process, MDP)이다. 마코프 특징을 직관적으로 설명하자면 현재가 주어졌을 때, 과거와 미래가 독립적임을 의미한다. 예를 들어서 내가 내일 얻을 시험 점수는 현재 내 상태와 오늘 내가 공부하는 양에만 의존함을 의미한다. MDP는 아래와 같이 크게 네 가지로 구성된다.

○ A set of states

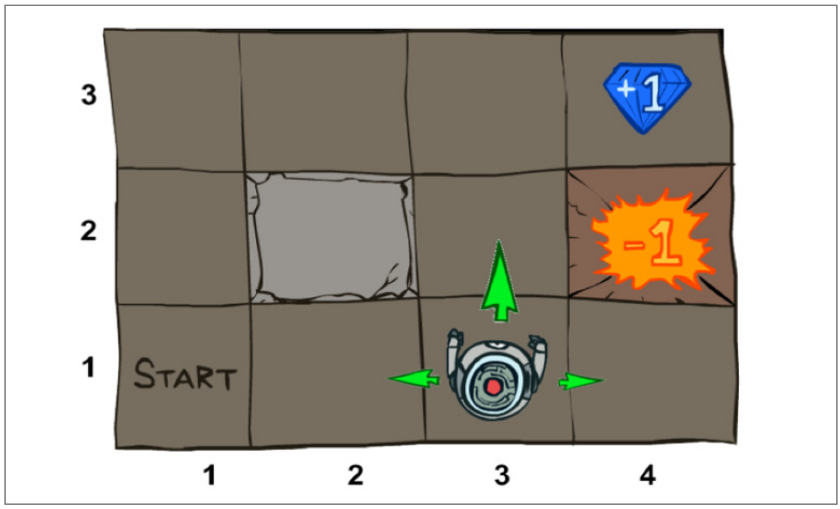
○ A set of actions

○ A transition function

○ A reward function

위의 네 가지를 아래의 예제를 통해 살펴보도록 한다. MDP를 설명하는 데 있어서 가장 흔하게 사용되는 예제는 아래와 같은 격자 공간 속에 로봇이 있는 상황이다.

[그림 2-33] 격자 공간 속에 로봇이 있는 상황



자료: Klein, D. & Abbeel, P. (2012). UC Berkeley CS188 Intro to AI-Course Materials(video).

[그림 2-33]에서 확인할 수 있듯이 로봇이 있을 수 있는 상태는 12개의 격자 중 하나이고, 이것이 상태 공간(state space)이다. 각 격자에서 로봇은 상하좌우로 이동하거나 제자리에 있을 수 있고, 이 다섯 가지의 행동이 행동 공간(action space)에 해당한다. 로봇의 격자 이동이 결정된 경우에는 원하는 방향으로 이동하게 되지만, 확률적인 경우에는 로봇이 위로 이동하려고 해도, 일정 확률로 오른쪽 혹은 왼쪽으로 이동하게 된다. 이렇듯 특정 상태에서 특정 행동을 했을 때, 다음번에 도달할 상태들의 확률을 나타낸 것이 전이 함수(transition function)이다. 마지막으로 강화학습에서 가장 중요한 보상 함수는 각 상태에 정의되는데, 위의 격자 공간에서 보석에 도달하면 +1의 보상을 , 불에 도달하면 -1의 보상을 얻게 된다. MDP가 주어졌을 때, 강화학습을 푼다는 것은 기대되는 미래 보상의 합을 최대로 하는 정책 함수를 찾는 것을 의미한다. 여기에는 여러 의미가 담겨있는데, 한 가지 중요한 점은 미래에 얻을 보상(reward)

이 포함되어 있다는 점이다. 현재 행동을 통해서 얻어지는 보상도 중요하지만, 궁극적으로는 현재뿐만 아니라 미래에 얻어지는 보상들도 다 고려를 해야 한다. 다음으로 고려해야 하는 점은 우리가 확률계(stochastic system)를 가정한다는 점이다. 현재 어떤 행동을 한다고 해서 원하는 상태에 도달하지 못할 수도 있게 되고, 이는 같은 보상을 얻을 수 있다면, 최대한 '빨리' 그 보상을 얻는 것이 더 좋음을 의미한다. 이러한 현상을 반영하는 것이 할인율(discount factor)이다. 이 값은 1보다 작은 값으로 설정되며, 시간이 지남에 따라서 얻어지는 보상에 이 값을 곱하게 된다.

그렇다면, MDP가 주어졌을 때, 최적의 정책 함수를 찾는 방법에 대해서 알아보고자 한다. 강화학습을 푸는 가장 기본적인 방법 두 가지는 값 반복(value iteration)과 정책 반복(policy iteration)이다. 이를 설명하기 위해서는 먼저 값(value)을 정의할 필요가 있다. 현재 즉각적으로 얻을 수 있는 보상뿐만 아니라, 해당 상태에서 시작했을 때 얻을 수 있는 보상들의 합의 기댓값을 알 수 있다면, 해당 함수를 매번 최대로 만드는 행동을 선택할 수 있을 것이고, 이렇게 함으로써 최적의 정책 함수를 구할 수 있게 된다. 바로 이 미래에 얻을 수 있는 보상들 합의 기댓값을 값 함수(value function), V(s)라고 부른다. 또한 비슷하게 현재 어떤 상태에서 어떤 행동을 했을 때, 미래에 얻을 수 있는 기대 보상을 행동값 함수(action value function) 혹은 Q function, Q(s,a)라고 부른다. 값 반복(value iteration)은 바로 이 값 함수를 구하는 방법을 지칭한다. 값 함수는 현재 상태뿐 아니라 미래의 상태들, 혹은 그 상태에서 얻을 수 있는 보상을 구해야 하기 때문에 직관적으로 정의하기 어렵다. 일반적으로 강화학습에서는 이 값 함수를 구하기 위해서 벨만 이퀘이션(Bellman equation)을 활용하고, 그 식은 다음과 같이 정의된다.

$$V(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V(s') \right]$$

위 수식은 좌변과 우변에 구하고 싶은 값 함수 V(s)가 들어가 있는 재귀식(recursive equation)임을 알 수 있고, V(s)를 제외한 나머지는 MDP와 같은 내용이다. V(s)를 임의의 값으로 초기화하고, 모든 상태 s에 대해서 위의 재귀식을 수렴할 때까지 수행하면서 항상 최적의 V(s)를 구할 수 있다.

앞에서 값 반복에 대해서 살펴보았다면 이번에는 값 반복과는 다른 정책 반복에 대해서 살펴보고자 한다. 정책 반복은 현재 가지고 있는 정책 함수의 성능을 평가하는 정책 평가와 이를 바탕으로 정책을 개선하는 정책 개선(policy improvement)의 단계로 이뤄져 있고, 이 두 단계를 정책 함수가 수렴할 때까지 번갈아가면서 수행한다. 일반적으로 정책 반복이 값 반복보다 빠르게 최적의 정책 함수로 수렴한다고 알려져 있다.

이제부터는 이렇게 MDP에서 모델이 주어지지 않았을 때 어떻게 최적의 정책 함수를 찾아내는지 알아보고자 한다. 이러한 문제를 일반적으로 모델 프리 강화학습(model-free reinforcement learning)이라고 부른다. 앞에서 설명했던 모델 기반 강화학습과의 가장 큰 차이는 더 이상 환경이 어떻게 동작되는지 알지 못한다는 점이다. 다시 말해서 주어진 상태에서 어떤 행동을 하고, '수동적으로' 환경을 알려주는 다음번 상태와 보상(reward)을 얻게 된다. 모델 프리 강화학습은 모델 기반 강화학습에 비해 몇 가지 구별되는 특징들이 있는데, 그 대표적인 것이 바로 탐사(exploration)다. 더 이상 환경이 어떻게 동작하는지 알지 못하기 때문에 직접 해보고 그 결과를 통해서 정책 함수를 점차 학습시켜야 한다. 이렇게 정의된 모델 프리 강화학습을 어떻게 풀 수 있는지 알아보고자 한다. 앞서 모델 기반 강화학습에 사용된 벨만 등식에서 T(s, a, s'), 바로 이 부분을 우리가 알 수 없기 때문에 벨만 이퀘이션을 직접 활용할 수는 없다.

정책 평가는 위에서 a를 π(s)로 치환해, 주어진 π를 평가하는 방법론이다.

$$V(s) = \max_a \sum_{s'} T(s,\pi(s),s') \left[R(s,\pi(s),s') + \gamma V(s')\right]$$

위의 수식이 모델 프리 강화학습에 사용되는 정책 평가 수식이다. T(s, π(s), s')는 모르는 값이지만, 만약 다음 상태(state)인 s'이 T라는 모델에서 나왔다면, 우리는 이 T의 합을 표본 평균(sample mean)으로 대체할 수 있게 된다. 이런 식으로 벨만 이퀘이션을 표본화로 대체하는 방법론 중 하나가 temporal difference(TD) 학습(learning)이다. MDP에서 경험은 내가 어떤 상태 s에서 주어진 정책 함수를 통해서 어떤 행동을 하고(a=π(s)), 그리고 그 결과로 다음번 상태 s'와 보상 r을 받는 것을 의미한다. 그리고 이런 (s ,a ,s',r)들로 이뤄진 경험들이 누적되었을 때, 이 데이터들을 바탕으로 값 함수(value function) V(s)와 행동 값 함수(action value function) Q(s,a)를 학습하게 된다. 벨만 이퀘이션을 Q(s, a)에 대해서 구한 식은 다음과 같다.

$$Q(s,a) \leftarrow \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q(s',a')\right]$$

위 수식을 이용해서 행동 값 함수 Q(s,a)를 업데이트 할 수 있다. 이 수식에서 T를 표본 추정(sample estimate)으로 교체할 수 있고, 이렇게 하면 다음과 같은 수식을 얻을 수 있게 된다.

$$Q(s,a) \leftarrow (1-\alpha) Q(s,a) + \alpha \left[R(s,a,s') + \gamma \max_{a'} Q(s',a')\right]$$

위 수식을 이용해서 강화학습 문제를 푸는 것을 Q러닝(Q-learning)이라고 한다. 이 수식을 이용해서 Q함수를 업데이트하기 위해서 필요한 것은 (s, a, s', r)들의 경험이다. 즉 어떤 상태에서 어떤 행동을 하고, 다음번 상태를 관측하고, 이때 얻어지는 보상까지 있으면 위의 수식을 이용해서 Q함수를 구할 수 있다. Q러닝이 가지는 장점은 모델을 모르는 상태에서도 최적 정책 함수(optimal policy function)를 구할 수 있다는 점이다. 그렇다고 해서 Q러닝이 절대 만능의 방법은 아니다. 모델 프리 강화학습의 가장 큰 단점은 탐사(exploration)이다. 강화학습 특징상 보상을 마지막에 한번, 혹은 띄엄띄엄 주는 경우가 있다. 이러한 탐사가 갖는 어려움은 상태 공간(state space)이 커짐에 따라서 더 부각된다.

이제부터 이러한 Q러닝을 딥러닝 기법과 결합한 방법인 딥 강화학습(deep reinforcement learning, DRL)에서 가장 중요한 것 중 하나가 탐사를 얼마나 잘하는가이다. DRL 중에 DeepMindtk에서 발표해 유명해진 DQN(Deep Q Network)이라는 방법론이 있다. 2016년 딥마인드의 알파고가 이세돌과 커제와의 대국에서도 승리할 수 있게 만들어준 알고리즘이 DQN이다. 기존의 강화학습의 알고리즘적 관점에서 DQN은 새로운 요소는 많이 없다. 하지만, Q러닝의 수식이 가지는 가장 큰 장점은 환경에 대한 그 어떤 정보 없이도, Q함수에 대한 업데이트를 할 수 있다는 점이다. 이러한 특성은 생각보다 큰 의미를 가진다. 왜냐하면 RL이 기존의 SL(지도학습)에 비해서 가지는 어려움이 기대되는 미래 보상의 합을 고려한다는 점인데, 위의 수식을 통해 단순히 하나의 경험만으로도 Q함수를 업데이트 시킬 수 있다는 것이다. 게다가 (s, a, s', r)에서 (s, a)를 고를 때 임의의 행동 a를 매번 고른다 하여도 무한한 시간이 흐른 뒤에 Q(s, a)는 항상 최적의 행동 값 함수(action value function)로 수렴한다는 점이 장점이다. 위 수식의 우변은 만약 현재 Q(s, a) 함수를 안다면

손쉽게 구할 수 있다. max a'Q(s', a')을 구하기 위해서는 가능한 행동 a의 수가 유한해야 한다. 다시 말해 모든 가능한 행동을 다 집어넣고, 제일 큰 Q값이 나오는 행동을 고르면 된다. 그리고 좌변의 Q(s, a)를 어떤 입력 (s, a)에 대한 Q 함수의 출력값이라 하고, 우변을 해당 입력에 대한 목적(target)이라고 본다면 위의 Q함수에 대한 벨만 이퀘이션은 우리가 지도학습에서 많이 사용하는 회귀 함수(regression function)에 대한 입출력 쌍을 만들어주는 것으로 해석할 수 있다.

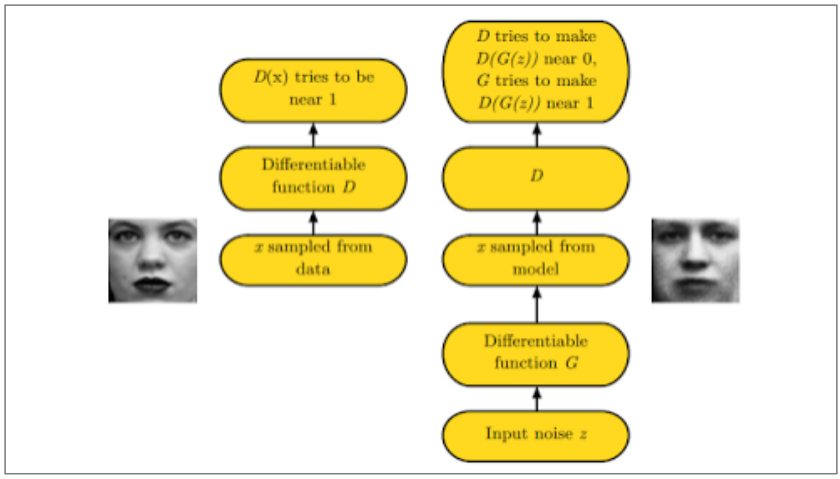
5. 딥러닝 기반의 기계학습 기법: GAN(Generative Adversarial Network)[8)]

NIPS 2016에서 가장 돋보인 키워드 중 하나는 GAN이라고 볼 수 있다. GAN(Generative Adversarial Network) 이름에서도 알 수 있듯이 'Adversarial' 이란 단어의 사전적 의미를 보면 '대립하는, 적대하는' 이라는 뜻이 있다. 대립하려면 최소 하나의 상대는 있어야 하니 GAN은 크게 두 부분으로 나누어져 있다는 것을 직관적으로도 알 수 있다. 만들어내는 파트(Generator)와 이렇게 만들어진 녀석을 평가하는 파트(Discriminator)가 있어서 서로 대립(Adversarial)하며 서로의 성능을 점차 개선해 나가자는 것이 GAN의 주요 개념이다. GAN을 좀 더 쉽게 설명하는 예시가 있는데 바로 다음 예와 같다. 지폐위조범(Generator)은 경찰을 최대한 열심히 속이려고 하고 다른 한편에서는 경찰(Discriminator)이 이렇게 위조된 지폐를 진짜와 감별하려고(Classify) 노력한다. 이런 경쟁 속에서 두 그룹 모두 속이고 구별하는 서로의 능력이 발전하게 되고 결과적으로는 진짜 지폐와 위조지폐를 구별할 수 없을 정도에 이른다는 것이다. 이러한 예시를 바탕으로 GAN을 좀 더 수학적

8) 유재준(2017)을 포괄적으로 인용하여 재구성함.

으로 설명하면 다음과 같다. Generative model G는 우리가 갖고 있는 data x의 distribution을 알아내려고 노력한다. 만약 G가 정확히 data distribution을 모사할 수 있다면 거기서 뽑은 sample은 완벽히 data와 구별할 수 없다. 한편 discriminator model D는 현재 자기가 보고 있는 sample이 training data에서 온 것인지 혹은 G로부터 만들어진 것인지를 구별하여 각각의 경우에 대한 확률을 estimate 한다.

[그림 2-34] GAN 예시



자료: Goodfellow, I., Pouget-Abadie, J., Mirza, M. (2014). Generative Adversarial Nets. NIPS 2016 Tutorial. Canada.

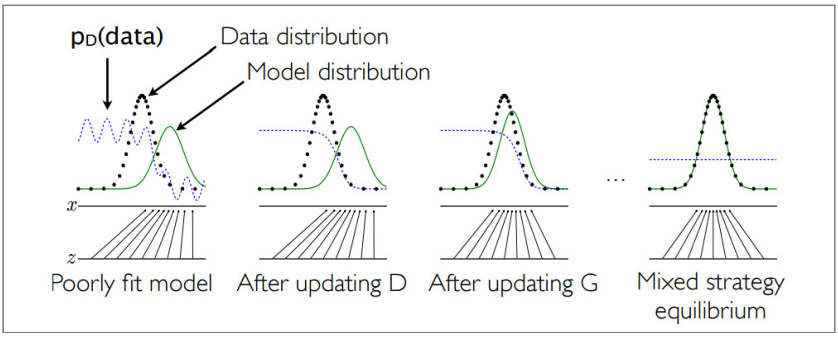
따라서 [그림 2-34]을 보면 알 수 있듯이 D의 입장에서는 data로부터 뽑은 sample x는 D(x)=1이 되고, G에 임의의 noise distribution으로부터 뽑은 input z를 넣고 만들어진 sample에 대해서는 D(G(z)) = 0이 되도록 노력한다. 즉, D는 실수할 확률을 낮추기(mini) 위해 노력하고 반대로 G는 D가 실수할 확률을 높이기(max) 위해 노력한다. 따라서 이 둘을 같이 놓고 보면 'minimax problem' 이라고 할 수 있다. *G*와 *D*가 둘 다 multilayer perceptrons model을 사용한다. Generator's dis-

tribution p_g over data x를 학습하기 위해 generator의 input으로 들어갈 noise variables $p_z(z)$에 대한 prior를 정의하고, data space로의 mapping을 $G(Z;\theta_g)$라 표현할 수 있다. 여기서의 G는 미분 가능한 함수로서 θ_g를 parameter로 갖는 multilayer perceptron이다. 한편 Discriminator 역시 multilayer perceptron으로 $D(x;\theta_d)$로 나타내며 output은 single scalar 값이 된다. $D(x)$는 x가 p_g가 아닌 data distribution으로부터 왔을 확률을 나타낸다. 따라서 이를 수식으로 정리하면 다음과 같은 value function $V(G,D)$에 대한 minimax problem을 푸는 것과 같아진다.

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

〔그림 2-35〕은 지금까지 위에서 설명한 내용과 수식들을 그려놓은 그림이다. 검은 점선이 data generating distribution, 파란 점선이 discriminator distribution, 녹색 선이 generative distribution이다. 또한 그림에서 위로 뻗은 화살표가 $x = G(z)$의 mapping을 보여준다. 그림을 살펴보면 처음 시작할 때는 p_g가 p_{data}와 전혀 다르게 생긴 것을 볼 수 있다. 이 상태에서 discriminator가 두 distribution을 구별하기 위해 학습을 하면 두 번째 그림과 같이 좀 더 잘 구별하는 distribution이 만들어진다. 이후 G가 현재 discriminator가 구별하기 어려운 방향으로 학습을 하면 세 번째 그림처럼 분포가 형성되고 이런 식으로 학습을 반복하다 보면 결국에는 $p_g = p_{data}$가 되어 discriminator가 둘을 전혀 구별하지 못하는 $D(x) = \frac{1}{2}$인 상태가 된다.

[그림 2-35] GAN 프로세스



자료: Goodfellow, I., Pouget-Abadie, J., Mirza, M. (2014). Generative Adversarial Nets. NIPS 2016 Tutorial. Canada.

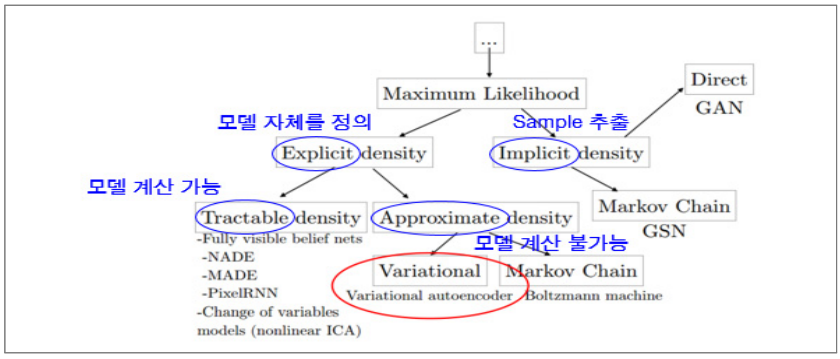
지금까지 NIPS 2014에서 발표되고 NIPS 2016에 가장 핫한 키워드 중 하나인 GAN에 대해서 이론적인 내용들과 예시들을 통해 알아보았다. 이미지처리 분야에서 GAN이 적용된 사례로는 사용자가 정확하게 스케치 하지 않아도 진짜 같은 그림을 생성해 주는 이미지 편집의 기능을 구현할 수 있게 해주었다. 또한 흐려진 이미지에 대해 '진짜 같아지는 학습'을 통해 더욱 선명한 이미지로 복원하는 기능을 제공하기도 했다. GAN의 등장으로 인해 인공지능은 '수동적 인식'에서 벗어나 '능동적 행동'을 할 수 있는 인공지능으로 발돋움했다. 이안 굿 펠로가 GAN을 발표한 지 2년 밖에 되지 않았지만, GAN은 이미지 생성부터 편집, 변환, 복원 등 다양한 애플리케이션에서 그 효용 가치를 보여주고 있다. 또한 지금은 이미지 뿐만 아니라 음성이나 자연어처리에도 GAN을 적용한 시도들이 활발히 이루어지고 있다.

6. 딥러닝 기반의 기계학습 기법: VAE(Variational Auto-Encoder)

VAE는 이름으로부터 이 기법의 특징을 어느 정도 유추해 볼 수 있다. Variational Auto-Encoder에서 variational 이라는 단어가 왜 붙게 되었는지부터 알아보고자 한다.

〔그림 2-36〕의 개념 구조도를 살펴보면 GAN과 VAE는 서로 Explicit density와 Implicity라는 갈래에서 갈라지게 된다. 이 두 갈래의 상위 개념을 보면 Maximum Likelihood(ML) 라는 개념이 존재한다. 즉, generative model이 하고자 하는 것은 ML의 원리를 바탕으로 학습을 해보자는 것으로 정리할 수 있다. 이때 어떤 식으로 likelihood를 다루느냐에 따라 다양한 전략들이 존재한다. 크게 나뉠 수 있는 방식이 위에서 언급한 Explicit density 방식과 Implicit density 방식이다.

〔그림 2-36〕 GAN과 VAE 구조도



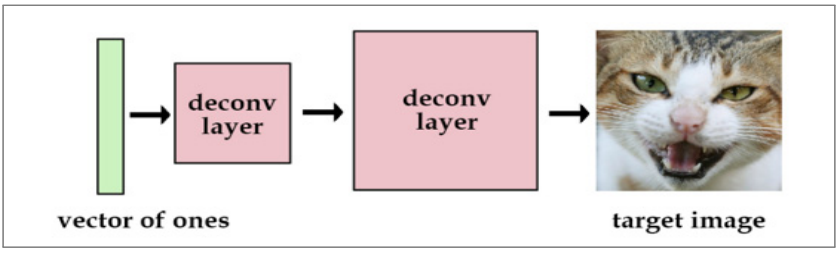
자료: Goodfellow, I., Pouget-Abadie, J., Mirza, M. (2014). Generative Adversarial Nets. NIPS 2016 Tutorial. Canada.

GAN 같은 경우 implicit density 방식에 속한다. 여기서 간접적인 방식이라 함은 보통 우리가 어떤 model에 대해 틀을 명확히 정의하는 대신 확률 분포를 알기 위해 sample을 뽑는 방법을 말한다.

반면에 explicit density 방식에 해당하는 학습 방법들은 모델을 명확히 정의하여 이를 최대화하는 전략을 택한다. 모델을 정의했기 때문에 다루기가 비교적 편하고 어느 정도 모델의 움직임이 예측 가능한 반면 당연히 우리가 아는 것 이상으로는 결과를 낼 수 없다는 한계가 있다. 이렇게 근사하는 방법에는 Monte Carlo approximations가 있다. 간단히 얘기하면 무작위로 많이 샘플을 뽑아서 분포를 유추해보자는 아이디어이다. 그러나 이러한 방식의 근사는 고차원을 다룰수록 잘 되지 않고 무엇보다 느리다는 단점이 있다. 이러한 점을 극복하기 위해서 VAE는 이와 달리 variational inference라는 보다 deterministic한 방식을 사용하여 density를 근사화한다.

VAE의 개념을 예시를 통해 조금 더 쉽게 알아보고자 한다. 다음과 같이 타깃 이미지를 고양이로 설정한 경우를 예로 들어보자.

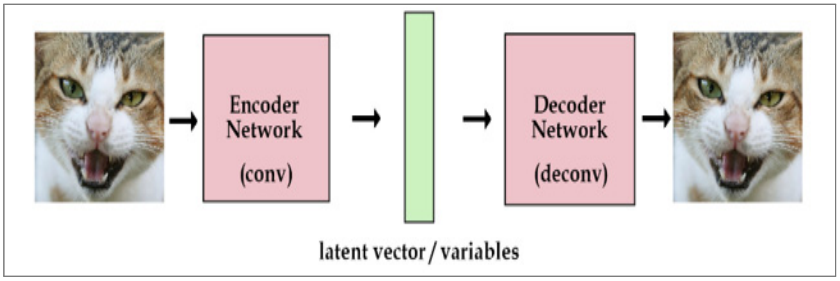
[그림 2-37] VAE 예시1



자료: Frans, K. (2016). Variational Autoencoders Explained. http://kvfrans.com/variational-autoencoders-explained/에서 2017. 11. 29. 인출.

어떤 vector를 입력으로 하여, deconvolution layer들을 거쳐서 이미지를 생성해낼 수 있도록 모델을 학습하고 싶어 하는 상황이 있다. 예를 들어, '[3.3 4.5 2.1 9.8] -> 고양이' , '[3.4 2.1 6.7 4.2]->개' 와 같은 모델을 학습하고 싶을 때, 입력 vector 값이 random이 아니라 target image에서 추출된 정보를 담고 있으면 좋겠다는 아이디어가 생긴다.

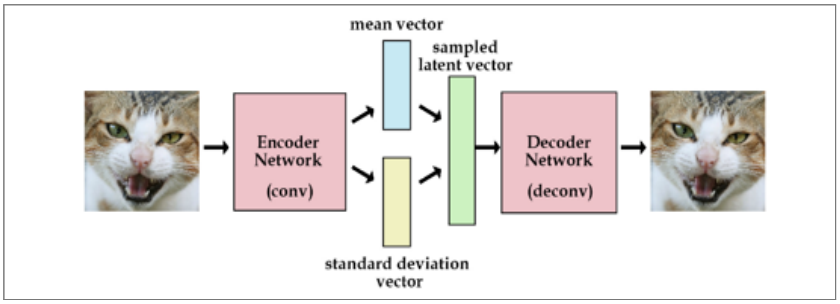
[그림 2-38] VAE 예시2



자료: Frans, K. (2016). Variational Autoencoders Explained. http://kvfrans.com/variational-autoencoders-explained/에서 2017. 11. 29. 인출.

입력 vector를 random 값이 아니라 이미지를 encoding한 값으로 대체하면, 다양한 이미지의 정보를 담는 것이 가능해진다. 이런 뒤 이미지의 정보를 encoding하여 latent variable에 저장하고, decoding하여 복원하는 Auto-Encoder 형태의 모델이 만들어진다. 최종적으로 원하는 형태는 이미지를 'memorize' 하는 것이 아니라 'generate' 하는 것이므로, latent variable에 어떤 constraint를 설정하여 constraint 내의 값의 변화에 따라 출력을 어느 정도 예측 가능하게 변화시킬 수 있는 모델을 만들려는 시도를 하게 된다. 이때 constraint는 기본적으로 Gaussian 모델을 사용한다.

[그림 2-39] VAE 예시3

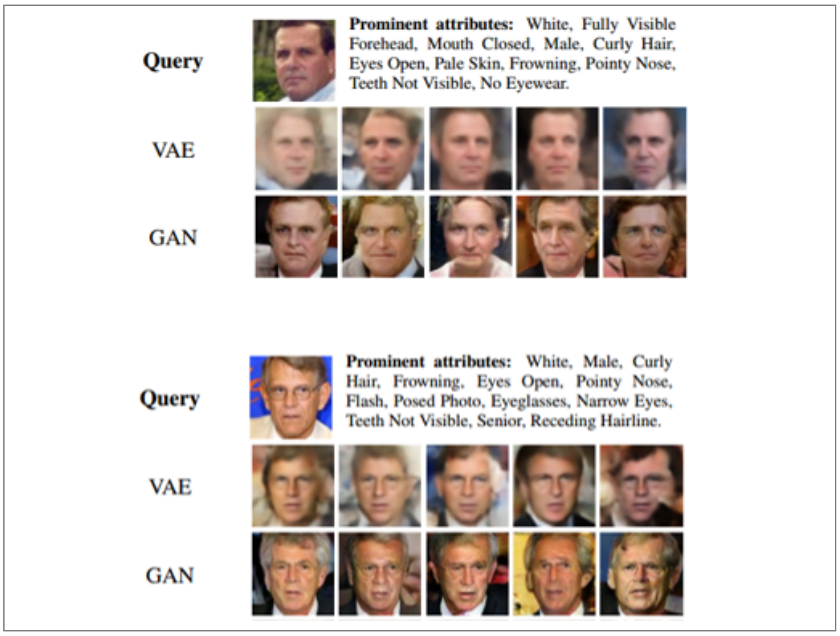


자료: Frans, K. (2016). Variational Autoencoders Explained. http://kvfrans.com/variational-autoencoders-explained/에서 2017. 11. 29. 인출.

직관적으로, '이미지를 정확하게 생성해 내는 것과 Gaussian 분포와 latent variable의 분포를 같게 만들어 주는 것'이 VAE의 목표가 되므로, cost function은 추정된 이미지와 실제 이미지 사이의 MSE와 latent variable의 분포와 Gaussian 사이의 KLD의 합으로 표현된다. 이때 KLD의 optimize를 위하여, encoding network가 Gaussian에서 추출된 real value를 생성해내는 것이 아니라 평균과 표준편차를 생성해 내는 reparameterization trick이 사용된다. 특정 constraint 즉, 평균과 표준편차를 가지는 Gaussian 내의 값들을 이용하여 이미지를 생성해 내는 것이 가능해지게 된다.

VAE를 앞에서 살펴본 GAN과 비교를 해보면 GAN 같은 경우 VAE에 비해 훨씬 명료한 값을 생성해내지만 VAE 같은 경우 GAN이 명료한 값을 생성해 내는 것에 비해 기존에 MSE를 cost function으로 사용하는 generative model의 문제점인 blurring 현상이 그대로 드러난다. 반면에 VAE는 원하는 특징을 출력에 반영하는 것이 가능하여, 출력하고자 하는 값을 출력에 나타낼 수 있다는 장점이 있다. 즉, original과 generated values의 직접 비교가 가능하다.

[그림 2-40] VAE와 GAN 출력 비교 사진



자료: Larsen, A. B. L., Sønderby, S. K., Larochelle, H., Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning, Vol(48), pp. 1558-1566. New York.

제 3 장

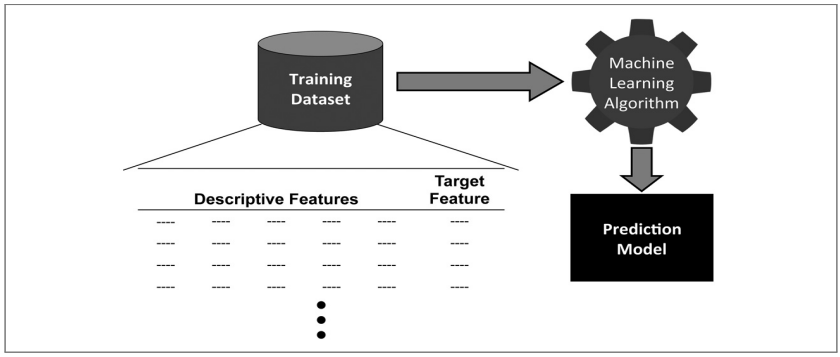
기계학습(Machine Learning) 통계기법 비교 연구

3 기계학습(Machine Learning) 통계기법 비교 연구

제1절 기계학습 통계기법 소개

여기에서는 기계학습 통계기법을 소개하고자 한다. 통계모형 활용 목적이 예측이기 때문에 지도학습(Supervised Learning)을 주로 살펴본다. 그동안 축적된 모형 구축을 위한 데이터셋(Training data)은 설명변수(Descriptive features)와 반응변수(Target feature)로 구성되어 있으며 이 데이터셋으로 기계학습 알고리즘을 활용하여 예측모형을 만든다. 이렇게 만들어진 예측모형을 새로운 데이터셋(Test data)에 적용하여 새로운 데이터에 예측값을 추정할 수 있다. 어떤 기법이 정확도가 높은지에 대한 부분은 모형 평가에서 다룰 예정이고 여기에서는 예측모형에 해당하는 기계학습 통계기법을 중심으로 설명하고자 한다.

[그림 3-1] 학습 모델



자료: Kelleher, J. D., Mac Namee, B. D'Arcy, A. (2015). Fundamental of machine learing for predictive data analytics-algorithms, worked examples and case studies. London: The MIT Press.

예측모형으로 활용할 수 있는 통계기법들은 요즘 화두가 되고 있는 딥러닝(Deep Learning)의 기반인 Neural Network뿐만 아니라, Regression, Shrinkage method, Decision Tree, Boosting, Random Forest, Support vector machines(SVM), Bagging, Deep Learning 등 매우 많다. 여기에서는 반응변수(종속변수)가 범주형인 경우 적용할 수 있는 Logistic regression, Shrinkage method인 Elastic Net, Decision Tree, Random Forest, Boosting, Support vector machines 방법에 대해 소개하고자 한다[9].

Logistic regression은 종속변수(y)가 범주형인 경우에 독립변수와의 관계를 모형화할 때 적용할 수 있는 방법으로, 연속이고 증가함수이고 [0,1] 사이에서 값을 갖는 $p(x)$에 대해

$$P(Y=1|x) = p(\beta_0 + \beta_1 x)$$

으로 수식화할 수 있다. 선형회귀모형에서는 x가 주어졌을 때 Y의 조건부 평균이지만, 로지스틱 회귀는 조건부확률을 연결함수(link function) p를 통해 모형화하는 것이다. 연결함수의 형태에 따라 $p(x) = \exp(x)/(1+\exp(x))$이면 로지스틱 모형, $p(x) = \exp(-\exp(x))$는 검벨(Gumbel) 모형, $p(x)$가 표준정규분포의 분포함수는 프로빗(probit) 모형이라고 하며 계산의 편리성으로 로지스틱 모형이 많이 사용된다.

단순 로지스틱 모형은

$$P(Y=1|X=x) = \frac{\exp(x^T\boldsymbol{\beta})}{1+\exp(x^T\boldsymbol{\beta})},\ (\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \ ..., \beta_p)^T)$$

9) 방법론 설명은 박창이, 김용대, 김진석, 송종우, 최호식(2011)의 내용을 포괄적으로 인용하여 재구성하였음

i.e. $\log(\frac{p(Y=1|X=x)}{p(Y=0|X=x)}) = x^T\boldsymbol{\beta}$ 이다.

$$\frac{P(Y=1|x+1)/P(Y=1|x)}{P(Y=0|x+1)/P(Y=0|x)} = \exp(\boldsymbol{\beta}_1)$$

을 오즈비(odds ratio)라고 한다. 오즈비는 x가 한 단위 증가할 때 $y=1$일 확률과 $y=0$일 확률 비의 증가율을 의미한다. 예를 들어 x는 소득이고 y는 어떤 제품에 대한 구입 여부(1=구입, 0=미구입)라 할 때, $\hat{\beta}_1 = 2$이면 소득이 한 단위 증가하면 물품을 구매하지 않을 확률에 대한 구매할 확률의 오즈비가 $\exp(2) = 7.38$배 증가하는 것으로 해석할 수 있다.

로지스틱 회귀에 대한 우도함수(likelihood function)는

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(\beta_0 + \beta_1 x_i)^{y_i}(1 - p(\beta_0 + \beta_1 x_i))^{1-y_i},$$

$$where \;\; p(x) = \exp(x)/(1+\exp(x))$$

이다. 로그 우도함수는 계수에 대한 비선형 함수이기 때문에 최대우도(maximum likelihood)추정치 $(\hat{\beta}_0, \hat{\beta}_1)$은 수치적 방법(numerical method)을 사용하여 구할 수 있다. 설명변수 x가 y를 설명하는 데 유의한지에 대한 유의성 검정은 우도비 검정 통계량

$$\chi^2 = -2\left(\max_{\beta_0} l(\beta_0, 0) - l(\widehat{\beta_0}, \widehat{\beta_1})\right)$$

이다. χ^2는 근사적으로 자유도가 1인 카이제곱 분포를 따르며 그 값이 크면 β_0가 0이 아니라고 결론을 내린다.

입력변수가 범주형인 경우 선형회귀와 같은 방법으로 가변수를 생성해서 분석할 수 있으며, 변수선택 역시 선형모형과 동일하게 적용할 수 있다. 선형회귀모형에서는 선택기준을 오차제곱합으로 사용하는 반면, 로지스틱 회귀에서는 로그 우도함수값을 사용한다. 로지스틱 회귀에서 변수선택을 전진선택법으로 하는 경우 각 단계마다 로그 우도함수값의 증가량을 구하고, 그 증가량이 가장 큰 변수부터 추가한다. 최종모형은 AIC나 BIC 등의 선택기준을 최소화하는 모형으로 선택한다.

로지스틱 회귀는 주어진 설명변수 x에 대해 반응변수 Y가 1이 될 확률 $P(Y=1|x)$를 추정하는데, 0과 1 사이의 합리적인 기준값 α를 절단값으로 선택하여 $\hat{P}(Y=1|X=x) > \alpha$ 이면 자료를 $Y=1$인 클래스로 분류하고$\hat{P}(Y=1|X=x) < \alpha$ 이면 자료를 $Y=0$인 클래스로 분류할 수 있다.

절단값 α를 결정할 때, 고려해야 하는 사항 첫 번째는 사전정보 고려이다. 사전정보에서 $y=1$인 자료가 상대적으로 많다면 절단값을 0.5보다 작은 값으로 고려할 수 있다. 두 번째로 적절한 손실함수를 고려해야 한다. $y=1$인 자료를 잘못 분류하는 손실이 $y=0$인 자료를 잘못 분류하는 손실에 비해 손실 정도가 심각하게 크다고 판단하는 경우 절단값 α를 작게 잡을 수 있다. 그 밖에도 전문가 의견이나 민감도, 특이도 등을 고려하여 α값을 결정할 수 있다.

로지스틱 회귀모형은 선형회귀모형과 마찬가지로 독립(설명)변수들 간에 강한 상관관계가 존재하는 다중공선성(Multicollinearity) 문제가 발생하는 경우, 회귀계수 추정량의 계산이 불가능할 수 있고, 추정량의 분

산이 커지는 등 신뢰할 수 없는 결과를 줄 수 있다. 다중공선성에 대한 문제를 해결하기 위해서는 강한 상관관계를 보이는 독립변수 중 일부를 사용하거나, 유사한 변수들을 주성분으로 묶는 PCA, 벌점화 방법론(method of penalization or regularization)인 능형회귀(Ridge regression), LASSO, Elastic Net 방법을 적용할 수 있다.

모형을 선정하는 데 있어서 근본적인 이슈 중 하나는 설명변수와 반응변수 간의 의존관계를 결정하는 모형의 복잡도를 어떻게 조절할 것인가이다. 통계학의 전통적 방법이라고 할 수 있는 모수적 추론(parametric inference)은 주어진 자료가 함수 분포를 따른다고 가정한다. 즉, 모형에 대한 가정을 한 뒤, 자료에서 모형의 모수에 대한 추론을 하는 방법이다. 그러나 모수적 추론 방식은 분포에 대한 가정을 하기 때문에 그 적용범위가 제한될 수밖에 없다. 최근 분석 대상이 되는 빅데이터(고차원 또는 대용량 자료)에 모수적 추론을 직접 적용하기에는 다중공선성 등의 현실적 어려움이 있다. 따라서 다양한 상황이 존재할 수 있는 빅데이터 분석에서는 모형에 대한 가정을 하지 않는 비모수적 추론(nonparametric inference)인 함수 추정(function estimation) 방법론을 적용할 수 있다. micro array 자료처럼 자료의 수보다 모수의 수가 더 크거나 무한차원의 함수형태인 모형의 경우 기존의 방법들로는 여러 문제를 야기할 수 있으며, 그 해결책으로 벌점화 방법론(method of penalization or regularization)을 고려하게 된다.

통계적 학습은 모형구축 데이터 $\{(x_i, y_i)\}_{i=1}^{n}$ 로부터 설명변수 $x \in \chi \subset R^p$와 반응(종속) 변수 $y \in Y$의 함수적 관계 f 를 찾는 다차원 공간상의 함수 추정의 문제로 볼 수 있으며, 예측력을 평가하는 기준으로 손실함수 $L(y, f(x))$를 사용한다. 그러나 주어진 자료는 유한개이고 추정할 함수는 무한 차원이므로 자료를 과대적합하는 해가 무수히 많다. 따라서 이 경우 경험위험을 최소화하는 해는 수학적으로 잘 정의되지 않는다.

Tikhonov와 Arsenin(1977)은 이러한 함수 추정 문제의 해가 수학적으로 잘 정의되도록 벌점화 방법을 제시하였다. 이는 벌점항 $J(f)$가 추가된 벌점화된 목적함수를 최적화한다.

$$\frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i) + \lambda J(f)$$

여기서 $\lambda > 0$은 훈련오차 $\frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i)$와 모형의 복잡도를 나타내는 $J(f)$ 간의 비중을 조절하는 모수이다.

Elastic Net은 변수선택과 추정을 동시에 할 수 있는 기법으로, 변수 간의 높은 상관관계가 있는 경우 적용할 수 있는 방법이다. Zou와 Hastie(2005)에 의해 제안된 Elastic Net 벌점화 기법은 $l_1 - norm$ 과 $l_2 - norm$의 볼록 결합(convex combination) 형태로 주어진다.

$$\lambda \sum_{j=1}^{p} ((1-\alpha)|\beta_j| + \alpha \beta_j^2), \alpha \in [0,1]$$

Elastic Net은 능형회귀(Ridge regression)와 Lasso 추정의 장점들을 포함한 방법으로, $l_1 - norm$ 과 $l_2 - norm$을 결합함으로써 상관관계가 있는 변수들을 모두 선택하여, 상관관계가 있는 변수들 중에서 하나의 변수만을 흔히 선택하는 Lasso 의 단점을 보완할 수 있다. 상관관계가 있는 변수들을 모두 선택하는 것을 그룹화 효과(grouping effect)라고 한다.

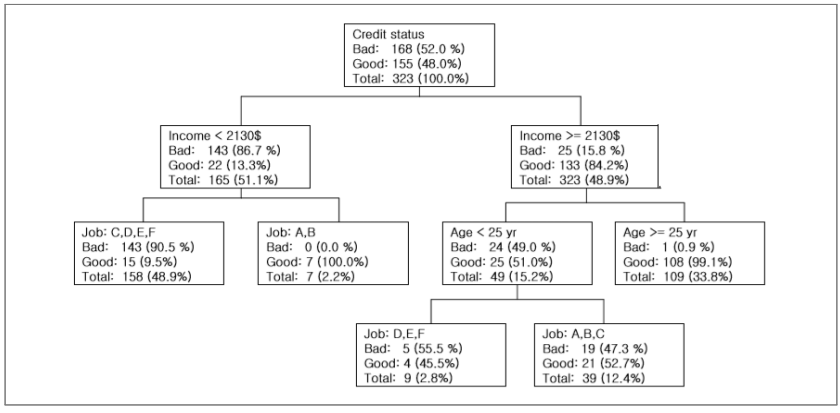
위 식에서 α, λ 값들이 주어졌을 때, 두 변수의 Elastic Net 추정계수를 각각 $\hat{\beta}_1(\alpha,\lambda)$, $\hat{\beta}_2(\alpha,\lambda)$ 라 하자. 어떤 상수 $M > 0$ 에 대해서 부등식

$$|\hat{\beta}_1(\alpha,\lambda) - \hat{\beta}_2(\alpha,\lambda)| < \frac{\sqrt{nM}}{\alpha\lambda}\sqrt{2(1-r_{12})}$$

이 성립한다. r_{12} 는 두 변수의 상관계수를 의미하며, 1에 가까울수록 두 계수의 차이가 0이 된다. 즉, 두 변수들 간에 상관관계가 크면 대응되는 추정치들의 값이 거의 동일하다고 할 수 있다.

지도학습 문제에서 최종모형의 예측력과 해석력이 중요한데, 상황에 따라 예측력보다 해석력을 더 중요하게 보는 경우가 있다. 장기요양등급 판정에서 심사 결과 등급 판정이 나온 경우 판정 대상자에게 해당 등급에 속한 이유를 설명할 수 있어야 하기 때문에 해석력이 더 중요시된다. 지도학습 방법 중 하나인 의사결정나무(Decision Tree)는 주어진 설명변수에 대해 반응 변수값을 예측하는 모형으로 분류나무(classification trees)와 회귀나무(regression trees) 모형이 있다. 의사결정나무는 결과를 나무 형태의 그래프로 표현할 수 있다.

[그림 3-2] 의사결정나무 예시



의사결정나무는 각 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성한다. 의사결정나무에 의해 생성된 규칙은 if - then으로 표현되어 이해하고 설명하기 쉽다.

의사결정나무의 형성과정은 크게 성장(growing), 가지치기(pruning), 타당성 평가, 해석 및 예측으로 이루어진다. 성장 단계는 각 마디에서 적절한 최적의 분리 규칙을 찾아서 나무를 성장시키는 과정으로서 적절한 정지규칙을 만족하면 중단한다. 가지치기 단계는 오차를 크게 할 위험이 높거나 부적절한 추론 규칙을 가지고 있는 가지 또는 불필요한 가지를 제거한다. 타당성 평가 단계에서는 이익도표(gain chart), 위험도표(risk chart), 평가자료(Test data)를 이용하여 의사결정나무를 평가하게 된다. 해석 및 예측 단계에서는 구축된 나무모형을 해석하고 예측모형을 설정한 후 예측에 적용한다.

의사결정나무는 다음과 같이 표현된다.

$$T(x;\Theta) = \sum_{m=1}^{M} \gamma_m I(x \in R_m)$$

여기서 R_m은 terminal node에서의 서로 배반인 설명변수 영역들을 나타내고, $\Theta = \mathrm{R}_m, \gamma_m$은 추정해야 할 모수를 의미한다. 전체 영역을 M개의 영역 R_1, ..., R_M으로 나누고 각 영역에서 상수값 γ_m으로 예측한다.

나무의 성장-분리 규칙에서 분리변수(split variable)가 연속형 변수인 경우 분리점을 c라고 정의하면, c보다 크면 오른쪽 R_1으로, 작으면 왼쪽 R_2로 나눈다. 범주형 변수는 전체 범주를 2개의 부분집합으로 나눈다.

분리 기준(split criteria)으로는 목표 변수의 분포를 얼마나 잘 구별하는가에 대한 측도로 순수도(purity) 또는 불순도(impurity)를 사용하며,

생성된 두 개의 자식마디에서 순수도의 합을 가장 크게 하거나 불순도의 합을 가장 작게 하는 분리변수를 분리기준으로 선택한다.

분류 나무에서의 불순도 측정 통계량은 카이제곱 통계량(chi-square statistics), 지니 지수(Gini index), 엔트로피 지수(Entropy index)가 있으며, 회귀 나무의 불순도 측정 통계량은 분산 분석에 의한 F-통계량(F-Statistics), 분산의 감소량이 있다.

앙상블(ensemble)은 주어진 자료로부터 여러 개의 예측모형을 만든 후 이러한 예측모형들을 결합하여 하나의 최종 예측모형을 만드는 방법을 의미한다. 최초로 제안된 앙상블 알고리즘은 Breiman(1996)의 배깅(bagging)으로, 부스팅(Boosting), 랜덤 포레스트(Random Forest)등의 방법이 많이 사용된다. 앙상블 방법은 예측력을 획기적으로 향상시킬 수 있음이 경험적으로 입증되었다.

부스팅의 기본 아이디어는 예측력이 약한, 랜덤하게 예측하는 것보다 약간 좋은 예측모형(weak learner)들을 결합하여 강한 예측모형(예측력이 최적에 가까운 예측모형)을 만드는 것이다.

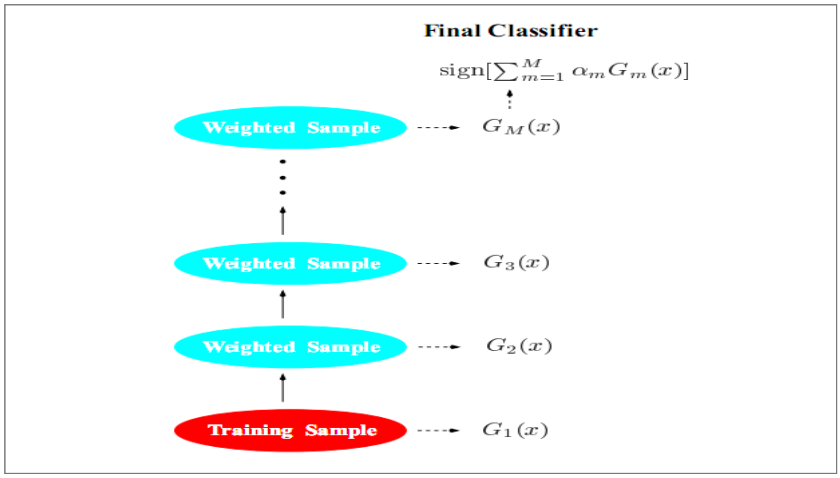
Freund와 Schapire(1997)에 의해 처음 제안된 부스팅 알고리즘은 분류(Classification) 문제를 해결하기 위한 방법으로 사용되었다. Weak classifier인 $G_m(x),\ i=1,2,\cdots,M$을 모아서 강력한 classifier인 $G(x)$을 만드는 방법은 다음과 같다.

$$G(x)=\text{sign}\left(\sum_{m=1}^{M}\alpha_m G_m(x)\right).$$

여기서, $G(x)$와 $G_m(x),\ i=1,2,\cdots,M$ 은 분류기(Classifier)들로 설명변수, x로부터 -1, 1로 가는 함수이다. $\alpha_m,\ i=1,2,\cdots,M$은 각 약한

분류기(weak classifier)들에 대한 가중치이다. 각각의 분류기(classifier)들은 가중표본(Weighted sample)으로부터 순차적으로 도출된다. 이때 α_m는 이전 단계에서 오분류한 표본에 대해 더 큰 가중치를 부여해서 다음 단계에서는 더 잘 분류할 수 있도록 업데이트된다. 이 알고리즘을 Adaboost 알고리즘이라 하고, 이 알고리즘을 도식적으로 표현하면 다음과 같다.

[그림 3-3] Adaboost 알고리즘 도식화



Adaboost 알고리즘의 정확한 가중치를 구하는 방법과 절차는 다음과 같다.

〔그림 3-4〕 Adaboost 알고리즘 스텝

1. Initialize the observation weights $w_i = 1/N, i = 1, \cdots, N$
2. For $m = 1$ to M
   (a) Fit a classifier $G_m(x)$ to the training data using weights w_i.
   (b) Compute
$$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$
   (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.
   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))], i = 1, \cdots, N$.
3. Output $G(x) = \text{sgn}[\sum_{m=1}^{M} \alpha_m G_m(x)]$.

위의 Adaboost 알고리즘은 G_m들이 이산 분류범주를 반환하기 때문에 discrete adaboost 알고리즘이라 한다(Friedman et al., 2000). 대신에 실수값을 반환하는 경우에 real adaboost 알고리즘을 사용할 수 있다.

Adaboost 알고리즘은 분류 문제에 적용 할 수 있는 알고리즘이기 때문에 회귀 문제에도 적용하기 위해서는 가법 모형(Additive model)을 고려하여 알고리즘을 일반화할 수 있다. 가법모형은 다음과 같은 수식으로 표현된다.

$$f(x) = \sum_{m=1}^{M} \beta_m b_m(x;\ \gamma_m).$$

여기서 $b(x;\gamma_m)$, $i = 1, 2, \cdots, M$은 basis function으로, 설명변수 x와 파라미터 γ로 반응변수를 예측하는 함수이다. 통상적으로는 벌점함수가 주어지면 모형구축 데이터(training data), $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$

에 대하여 아래 식과 같이 그 평균값을 최소로 하는 모수값을 구하는 것이 보통이다.

$$\min_{\beta_m, \gamma_m} \sum_{i=1}^{N} L\left(y_i, \sum_{m=1}^{M} \beta_m b_m(x_i; \gamma_m)\right).$$

이런 식의 해를 구하는 것은 계산이 굉장히 복잡하고 어렵지만 하나의 basis function을 최소화하는 문제는 상대적으로 쉽기 때문에 순차적인 방법으로 최소화 문제를 근사시킬 수 있다. 이 알고리즘을 Forward stagewise additive modeling이라 한다.

[그림 3-5] Forward stagewise additive modeling 스텝

1. Initialize $f_0(x) = 0$.
2. For $m = 1$ to M :
   (a) Compute
   $$(\beta_m, \gamma_m) = \underset{\beta, \gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)).$$
   (b) Set $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$

discrete Adaboost 알고리즘은 다음의 벌점함수를 사용한 forward stagewise additive modeling이라고 할 수 있다.

$$L(y, f(x)) = \exp(-yf(x))$$

위의 벌점함수를 지수벌점함수(Exponential loss function)라 부른다. 또한, 다음 식을 어렵지 않게 증명할 수 있는데,

$$f^*(x) = \arg\min_{f(x)} \boldsymbol{E}_{Y|x}(e^{-Yf(x)}) = \frac{1}{2}\log\frac{\Pr(Y=1|x)}{\Pr(Y=-1|x)}$$

$$\text{즉, } \Pr(Y=1|x) = \frac{1}{1+e^{-2f^*(x)}}.$$

위의 결과는 지수 벌점함수가 일반적인 로지스틱회귀(logistic regression) 문제를 다루는 자연스러운 벌점 함수임을 의미한다.

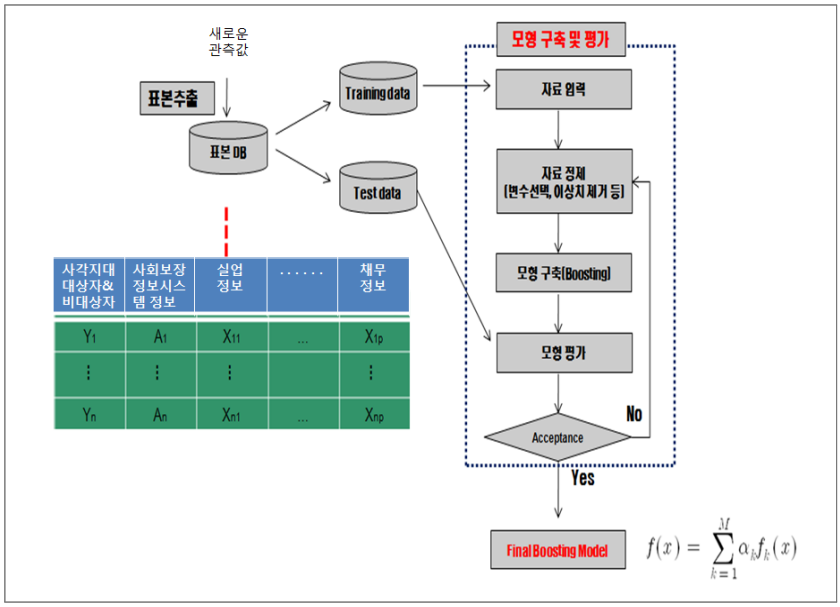
부스팅 알고리즘에서 basis function으로 주로 사용되는 것은 Stump이다. Stump는 가장 간단한 형태의 의사결정나무(Decision tree)로 root node에서 한번만 분기하는 즉 terminal node가 두 개인 것을 말한다. Friedman(2001)에 의해 제안된 Gradient tree boosting 알고리즘 수행 절차는 다음과 같다.

[그림 3-6] Gradient tree boosting 알고리즘 스텝

1. Initialize $f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to M :
   (a) For $i = 1, 2, \cdots, N$ Compute
   $$\gamma_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal resions R_{jm}, $j = 1, 2, \cdots, J_m$
   (c)
   $$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
3. Output $\hat{f}(x) = f_M(x)$

모형을 구축한 후에는 최적의 모형을 선택하고 평가하는 작업이 수반된다. 구축된 여러 모형들 중에서 오차가 가장 작은 모형을 선택한 후에 선택된 모형에 대해서 최종적으로 평가하는 과정을 거친다.

〔그림 3-7〕 최종 boosting 모형 구축 과정



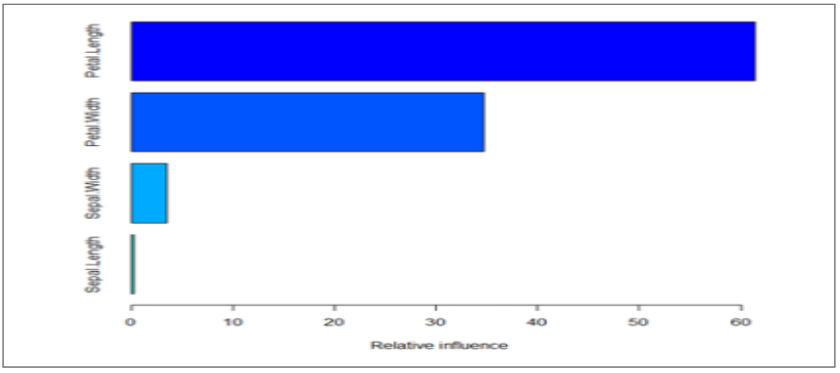
Friedman(2001)은 변수의 상대적 중요도를 구하였는데, 변수의 중요도는 통계모델에서 특정 변수가 반응변수에 미치는 영향에 대한 연관성 크기를 나타낸다. 트리 기반 모형에서 x_j 변수에 대한 상대적 영향도는

$$\hat{J}_j^2 = \sum_{splits\ on\ x_j} I_t^2$$

으로 표현할 수 있고, I_t^2는 변수 x_j 에 의해 순수도가 높아졌는지를 의미

한다. 부스팅 모형에서는 부스팅 알고리즘에 의해 생성된 모든 트리에 대해 x_j 변수의 상대적 영향 평균값을 x_j 변수의 상대적 영향도라고 말한다.

[그림 3-8] 상대적 영향도 예시



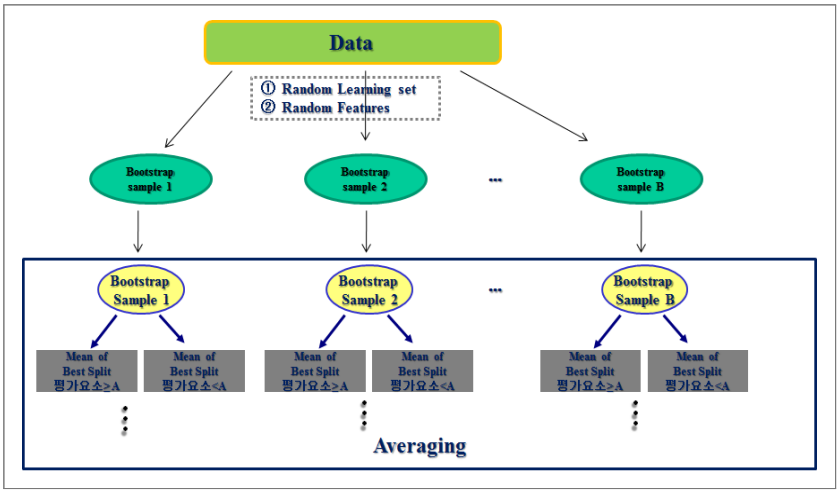
랜덤 포레스트(Random Forest)는 무작위성을 최대로 주기 위하여 부트스트랩(주어진 자료로부터 동일한 크기의 표본을 랜덤 복원 추출)과 더불어 독립변수들에 대한 무작위 추출을 결합한다. 서로 연관성이 약한 learner를 여러 개 만들어내는 기법으로 알고리즘 순서는 다음과 같다.

우선, 모형구축 데이터(Training data) $L = \{y_i, x_i\}_{i=1}^{n}$, $x_i \in R^p$ 에 대해 n개의 자료를 이용한 부트스트랩 표본 $L^* = \{y_i^*, x_i^*\}_{i=1}^{n}$ 을 생성한다. 그다음 L^*에서 독립변수들 중 $k(k \ll p)$개만 무작위로 뽑아 의사결정나무를 생성한다. 이때, 의사결정나무는 정해놓은 s 단계까지 진행한다. 위 두 단계를 여러 번 반복하여 생성된 의사결정나무들을 선형 결합하여 최종 learner를 만든다.

랜덤 포레스트 알고리즘에서 고려해야 하는 사항으로 부트스트랩 표본을 몇 개 생성할 것인지, k/d값을 어떻게 할 것인지, 선형결합의 형식을

어떻게 할 것인지에 대한 여러 가지 선택의 문제가 존재한다. 많은 경우 부트스트랩 표본의 개수는 지나치게 적게 하지는 않고, 선형결합의 형식은 각각의 의사결정나무의 결과들에 대하여 회귀분석에서는 평균을, 분류문제에서는 다수결원칙을 적용하는 방식을 사용한다.

[그림 3-9] 랜덤 포레스트



일반적으로 분류문제에 있어서 0-1 손실함수를 이용하지만, 랜덤 포레스트에 대한 이론적 설명에는 마진함수(margin function)와 그에 기반을 둔 랜덤 포레스트 강도(strength)를 이용한다.

먼저 랜덤 포레스트에 사용된 의사결정나무의 모집단을 $\mathrm{F}(\Theta)$로, 이 모집단의 의사결정나무는 $\{f(\bullet, \Theta_k), k=1,...,K\}$ 로 표기한다. 만약 (x,y)가 (X, Y)의 분포로부터 생성된 표본이고 각각의 분류함수 $f(\bullet, \Theta_k)$가 투표방식으로 분류한다면,

$$\frac{1}{K}\sum_{k=1}^{K} I(f(x,\Theta_k)=y) - \arg max_{l\neq y}\left(\frac{1}{K}\sum_{k=1}^{K} I(f(x,\Theta_k)=l)\right) < 0$$

이면 잘못된 분류를 하게 될 것이다.

이 사실을 이용하여 랜덤 포레스트 F(Θ)의 마진함수, 예측오차, 강도를 다음과 같이 정의할 수 있다.

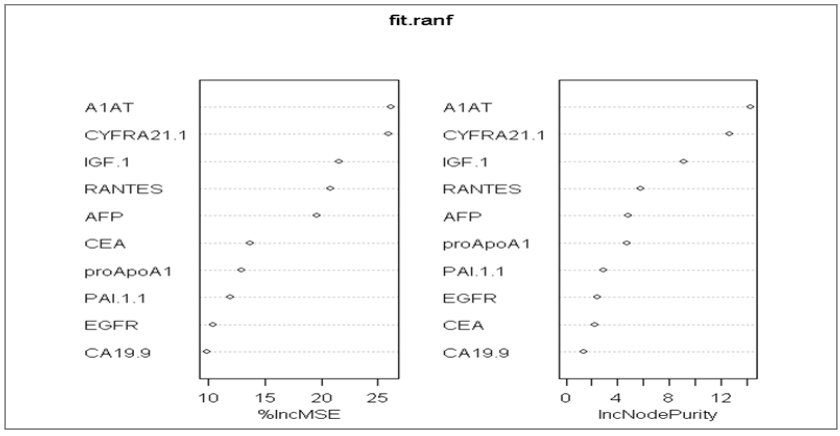
$$M_{F(\Theta)}(x,y) = P_{\Theta}(f(x,\Theta)=y) - \max_{l\neq y} P_{\Theta}(f(x,\Theta)=l),$$

$$R_{F(\Theta)} = P_{X,Y}(M_{F(\Theta)}(X,Y) < 0),$$

$$S_{F(\Theta)} = E_{X,Y}[M_{F(\Theta)}(X,Y)].$$

랜덤 포레스트에서는 변수에 대한 상대적 중요도 지수를 제공하는데, 특정 변수에 대한 중요도 지수는 그 변수를 포함하지 않고 모형을 구축하였을 경우에 어느 정도 예측오차가 줄어드는지를 보여주는 것이다.

[그림 3-10] 랜덤 포레스트의 상대적 중요도 예시

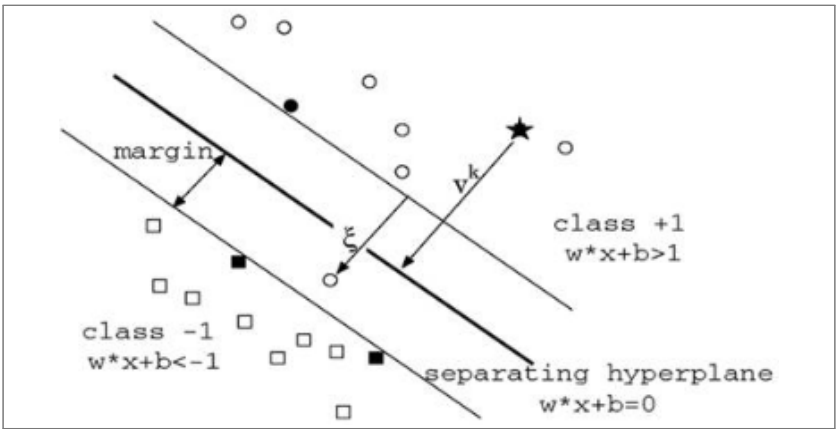


서포트 벡터 기계(Support vector machines)는 Cortes와 Vapnik (1995)가 제안하였다. 서포트 벡터 기계는 서포트 벡터 분류를 의미하는 것이지만 분류문제뿐만 아니라 회귀문제에도 적용 가능하며, 예측이 정확하고 여러 가지 형태의 자료에 대해서 적용하기 쉽기 때문에 최근 여러 가지 예측 문제에서 많이 쓰이고 있다. 기존의 대표적인 분류 방법인 로지스틱 회귀는 설명변수가 주어졌을 때 반응변수에 대한 조건부 확률을 추정하는 데 비해, 서포트 벡터 기계는 확률 추정을 하지 않고 직접 분류결과에 대한 예측만을 한다. 따라서 분류 효율 자체만 보면 서포트 벡터 기계가 로지스틱 회귀나 판별분석 등의 확률 추정방법들에 비해 예측력이 전반적으로 높다. 그러나 확률 추정값을 요구하는 적용분야에서는 잘 사용되지 않는다. 서포트 벡터 기계는 벌점화 기법의 하나로 볼 수 있고, 계산량이 자료의 수에 비례하지만 차원의 크기에는 덜 민감하므로 자료의 개수보다 차원이 더 큰 ($n \ll p$) 경우에 장점이 있다.

선형 서포트 벡터 기계는 모형구축 데이터를 이용하여 $f(x) = < w,x > + b$로 정의되는 선형 분류함수를 얻은 후 $f(x)$의 부호를 입력하여 새로운 설명변수 x에 대한 반응 변수값을 예측한다. 여기에서 $< w, x > = w^T x$ 이다. 자료가 선형함수로 분리 가능하므로 모든 $i(=1, ..., n)$에 대하여 $y_i f(x_i) > 0$인 선형 초평면(linear hyperplane) $f(x) = 0$가 무수히 많이 존재한다. 서포트 벡터 기계는 다음과 같이 모형구축 데이터들의 마진(margin) $C > 0$를 최대화하는 유일한 초평면을 찾아준다.

제약조건 $y_i(< w, x_i + b >) \geq C,\ i = 1,...,n$에 대하여
$\max_{b,w:\ \| w \| = 1} C$

[그림 3-11] SVM의 선형 분류



자료: WordPress. (2009). https://onionesquereality.wordpress.com/2009/03/22/why-are-support-vectors- machines-called-so/에서 2017. 11. 29. 인출.

자료가 선형적으로 분리 가능하지 않은 경우(linearly nonseparable case)에는 슬랙 변수(slack variable) $\xi_1,...,\xi_n$을 도입함으로써 어느 정도 오분류를 허용하며 마진 C를 최대화하는 방식을 취한다. 즉, 제약조건 $y_i(< w,x_i > + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,...,n$하에서

$$\max_{b,w:\ \| w \| = 1}(\frac{1}{2} \| w \|^2 + C\sum_{i=1}^{n}\xi_i).$$

이때, $C > 0$은 모형구축오차와 벌점 간의 상대적 크기를 조절하는 벌점 모수로 볼 수 있다.

제2절 기계학습 통계기법 장·단점 비교 분석

로지스틱 회귀의 가장 큰 장점은 모형이 안정적이라는 것이다. 또한 계수값으로 해석이 쉬우면 계산이 어렵지 않다. 로지스틱 회귀의 단점은 오직 linear decision boundary를 제공한다는 점이며 교호작용(Interaction term)을 모형에서 자동으로 고려해주지 않아, 두 변수 간의 교호작용 효과를 고려하려면 변수를 생성해야 한다. 또한, 독립변수의 변환에 invariant 하지 않다.

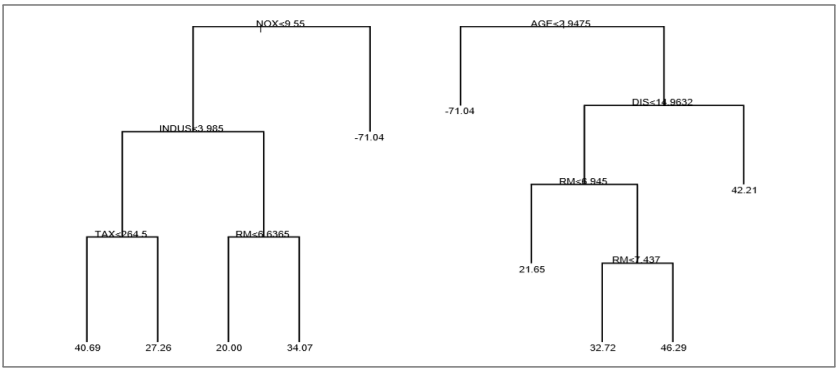
의사결정나무는 if-then 형식의 이해하기 쉬운 규칙을 생성하며, 분류작업이 용이하고, 연속형 변수와 범주형 변수를 모두 취급할 수 있으며, 모형에 대한 가정(선형회귀의 경우 등분산성, 선형성 등)이 필요없는 비모수적 방법이라는 장점이 있다. 또한 가장 설명력이 있는 변수에 대하여 최초로 분리가 일어나는 특징을 가진다. 그리고 이상치(outlier)에 덜 민감하다는 장점도 있다.

의사결정나무의 단점은 반응변수가 연속형인 회귀모형에서는 그 예측력이 떨어진다. 일반적으로 복잡한 나무모형은 예측력이 저하되고 해석이 어려우며, 상황에 따라 계산량이 많을 수도 있으며, 베이즈 분류경계가 사각형(recangle)이 아닌 경우에는 좋지 않은 결과를 줄 수 있다는 것이다. 특히, 자료에 약간의 변화가 있는 경우에 전혀 다른 결과를 줄 수도 있는(분산이 매우 큰) 불안정한 방법이다.

노드 사이즈 5인 Boston Housing data 의 두 부트스트랩(주어진 자료로부터 동일한 크기의 표본을 랜덤 복원 추출) 샘플들의 트리를 비교해 보면 전혀 다른 모형이 나오게 된다. 의사결정나무가 불안정한 이유는 첫 번째 노드에서 분리변수를 찾을 때 서로 비슷한 예측력을 갖는 변수가 다수 존재하기 때문이다. 따라서 자료의 작은 변화에도 첫 번째 노드의 분

리변수가 바뀔 수 있다. 이는 첫 번째 분리변수가 바뀌면 자식노드에 포함되는 자료가 완전히 바뀌게 되고 최종 의사결정나무가 완전히 달라짐을 의미한다.

[그림 3-12] Decision Tree의 불안정성 예시



부스팅 알고리즘은 매우 높은 예측력을 가지고 있고 독립변수의 변환에 invariant하며, Interaction terms을 고려할 수 있어서 비선형 효과까지 고려할 수 있다.

부스팅의 단점은 모형에서 Parameter tunning 작업이 까다롭다고 할 수 있다.

랜덤 포레스트의 장점은 부스팅과 마찬가지로 매우 높은 예측력을 가지고 있고, 독립변수의 수가 많을 때에도 좋은 결과를 얻을 수 있다. 또한, 이상치에 둔감하며 독립변수의 변환에 invariant하다. Interaction terms를 반영하여 비선형 효과도 고려할 수 있다.

랜덤 포레스트의 단점은 이론적 설명이 부족하며 최종 결과에 대한 해석이 어렵다.

Support Vector Machine은 다차원 공간을 초평면으로 경계를 분류

하는 방법으로 예측력이 좋아 기업체에서 많이 적용하고 있는 모형이다. 분류 및 회귀 문제에서 적용 가능하며, 데이터에 잡음(noise)이 있어도 크게 영향받지 않고 과적합(overfitting)되지 않는다. 그리고 높은 예측력도 SVM의 장점 중 하나이다.

SVM의 단점으로는 모형구축을 위한 데이터셋(training data)의 샘플 수와 차원(dimension)이 크다면 속도가 느려질 수 있다. 그리고 최적의 모형을 찾기 위해 커널과 모형에서의 Tuning parameter 설정이 중요하다. 또한, 모형에 대한 해석도 쉽지 않다.

〈표 3-1〉 A summary of models and some of their characteristics

Model	Allow n 〈 p	Pre-processing	Interpretable	Automatic feature selection	# Tuning parameters	Robust to predictor noise	Computation time
Linear regression[1)]	×	CS, NZV, Corr	✓	×	0	×	✓
Elastic net/lasso	✓	CS, NZV	✓	✓	1-2	×	✓
Support vector machines	✓	CS	×	×	1-3	×	×
Single trees	✓		○	✓	1	✓	✓
Model trees/rules[1)]	✓		○	✓	1-2	✓	✓
Random forest	✓		×	○	0-1	✓	×
Boosted trees	✓		×	✓	3	✓	×
Logistic regression[2)]	×	CS, NZV, Corr	✓	×	0	×	✓

주: 1) regression only.
2) classification only.
Symbols represent affirmative (✓), negative (×), and somewhere in between (○).
CS = centering and scaling.
NZV = remove near-zero predictors.
Corr = remove highly correlated predictors.
자료: Kuhn, M. & Johnson, K. (2013). Applied Predictive Modeling. USA: Springer Science & Business Media.

〈표 3-2〉 Classification: Predicting a Categorical Target Variable

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Cases
Decision Trees	Partitions the data into smaller subsets where each subset contains (mostly) responses of one class (either “yes” or “no”)	A set of rules to partition a data set based on the values of the different predictors.	No restrictions on variable type for predictors.	The label cannot be numeric. It must be categorical.	Intuitive to explain to nontechnical business users. Normalizing predictors is not necessary.	Tends to overfit the data. Small changes in input data can yield substantially different trees. Selecting the right parameters can be challenging.	Marketing segmentation, fraud detection.
Artificial Neural Networks	A computational and mathematical model inspired by the biological nervous system. The weights in the network learn to reduce the error between actual and prediction.	A network topology of layers and weights to process input data.	All attributes should be numeric.	Prediction of target (label) variable, which is categorical.	Good at modeling nonlinear relationships. Fast response time in deployment.	No easy way to explain the inner working of the model. Requires preprocessing data. Cannot handle missing attributes.	Image recognition, fraud detection, quick response time applications.

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Cases
Support Vector Machines	Essentially a boundary detection algorithm that identifies/ defines multidimensional boundaries separating data points belonging to different classes.	The model is a vector equation that allows us to classify new data points into different regions (classes).	All attributes should be numeric.	Prediction of target (label) variable, which can be categorical or numeric.	Very robust against overfitting. Small changes to input data do not affect boundary and thus do not yield different results. Good at handling nonlinear relationships.	Computational performance during training phase can be slow. This may be compounded by the effort needed to optimize parameter combinations.	Optical character recognition, fraud detection, modeling "black-swan" events.
Ensemble Models	Leverages wisdom of the crowd. Employs a number of independent models to make a prediction and aggregates the final prediction.	A metamodel with individual base models and a aggregator.	Superset of restrictions from the base model used.	Prediction for all class values with a winning class.	Reduces the generalization error. Takes different search space into consideration	Achieving model independence is tricky. Difficult to explain the inner working of the model.	Most of the practical classifiers are ensemble.

자료: Vijay Kotu & Bala Deshpande. (2014). Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner. Morgan Kaufmann.

〈표 3-3〉 Regression : Predicting a Numeric Target Variable

Algorithm	Description	Model	Input	Output	Pros	Cons	Use Cases
Linear Regression	The classical predictive model that expresses the relationship between inputs and an output parameter in the form of an equation.	The model consists of coefficients for each input predictor and their statistical significance. A bias (intercept) may be optional.	All attributes should be numeric.	The label may be numeric or binominal.	The workhouse of most predictive modeling techniques. Easy to use and explain to nontechnical business users.	Cannot handle missing data. Categorical data are not directly usable, but require transformation into numeric.	Pretty much any scenario that requires predicting a continuous numeric value.
Logistic Regression	Technically, this is a classification method. But structurally it is similar to linear regression.	The model consists of coefficients for each input predictor that relate to the "logit." Transforming the logit into probabilities of occurrence (of each class) completes the model.	All attributes should be numeric.	The label may only be binominal.	One of the most common classification methods. Computationally efficient.	Cannot handle missing data. Not very intuitive when dealing with a large number of predictors.	Marketing scenarios (e.g., will click or not click), any general twoclass problem.

자료: Kotu, V. & Deshpande, B. (2014). Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner. USA: Morgan Kaufmann.

제 4 장

기계학습(Machine Learning) 기반 예측모형 평가방법 연구

4 기계학습(Machine Learning) 기반 예측모형 평가방법 연구

제1절 모형 평가방법 개념 정의

모형 평가방법에 대한 개념 정의를 하기에 앞서 생각해볼 수 있는 것은 좋은 모형의 조건(what are good models?)이다.

좋은 모형의 조건은 간단(simple)하고 타당(valid)하며, 강건(robust)함을 요소로 가지고 있어야 한다(Sayama, 2015).

모형의 단순함(simplicity)는 무엇을 모델링할 것인지의 핵심요소이다. 우리가 모형을 구축하려는 주된 이유는 현실을 짧고 단순하게 묘사하고 싶기 때문이다. Occam's razor 의 유명한 원칙에 의하면, 두 모형이 같은 예측력을 가지고 있을 때, 간단한 모형을 선택해야 한다는 것이다. 이것은 이론이나 논리적으로 입증된 사실이 아니라 과학에서 일반적으로 받아들여지는 관행이라고 할 수 있다. 간소화(law of parsimony)는 경제성과 통찰력 측면에서 좋다. 모형에서 손실(loss)이 없다면 모수(parameters), 변수(variables), 가정(assumption)을 하나라도 제거하는 것이 좋다.

모형의 타당성(validity)은 모형의 예측이 관측된 현실과 얼마나 가깝게 일치하는가이다. 모형의 타당성은 실용적인 관점에서 가장 중요하다. 만약 모형의 예측이 합리적으로 관측치와 일치하지 않는다면 모형은 현실을 나타내지 못하는 것이고, 쓸모없게 된다. 따라서 모형의 예측력뿐만 아니라 사용하는 가정에 대한 타당성을 확인하는 것이 중요하다. 즉, 기존의 지식과 상식을 고려하여 모형에서 사용된 각각의 가정이 의미가 있는지를 살펴보아야 한다.

때로는 모형의 단순성과 타당성 사이에 상충관계(trade off)가 존재한다. 모형의 복잡도를 증가시키면 모형은 관측된 데이터를 잘 설명하지만 단순성에서 멀어질 수 있으며 과적합(overfitting)의 문제도 생길 수 있다. 따라서 모형을 설계할 때 두 기준 사이에서 균형을 맞추는 것이 필요하다.

마지막으로, 모형의 강건성(robustness)은 모형의 예측이 모형의 가정(assumptions)과 모수(parameters) 세팅의 사소한 변화에 얼마나 둔감한지를 나타낸다. 모형의 강건성은 현실세계로부터 가정을 추가하거나 모수값을 측정할 때 항상 오류(error)가 존재하기 때문에 중요한 부분이다. 만약 모형의 예측력이 미세한 변화에도 민감하다면, 그것으로부터 파생된 결론은 아마도 신뢰할 수 없을 것이다. 하지만 모형이 강건하다면, 결과는 가정과 모수의 작은 변화에도 유지될 것이고, 현실에 더 많이 적용될 수 있고 신뢰를 가질 수 있다.

위의 좋은 모형이 갖추어야 할 요소들을 고려해볼 때, 모형 평가는 하나의 자료를 분석하는 경우 여러 통계모형들을 비교하여 분석하는 것이 바람직하며, 이 중 자료를 최적으로 설명하는 모형을 선택하는 것이 좋다. 최적의 모형을 선택하기 위해서는 여러 모형을 비교 및 평가해야 하고, 선정된 최적의 모형이 다른 모형에 비하여 우수하다는 사실을 입증해야 한다.

모형을 평가하는 방법에서 고려해야 할 사항은 다음 4가지이다.

○ 예측력: 얼마나 예측을 잘하는가?

○ 해석력: 입력변수와 출력변수와의 관계를 잘 설명하는가?

○ 효율성: 얼마나 적은 수의 입력변수로 모형을 구축했는가?

○ 안정성: 모집단 내 다른 자료에 적용하는 경우 같은 결과를 주는가?

예측력, 해석력, 효율성, 안정성 모두 모형을 평가할 때 중요한 요소이지만 예측 문제에서 가장 중요한 고려 사항은 예측력이다. 아무리 안정적이고 효율적이며 해석이 쉬워도, 실제 문제에 적용했을 경우 빗나간 결과를 도출하는 경우, 모형을 적용하는 의미가 없다. 따라서 모형의 평가란, 예측(prediction)을 위해 만든 모형이 임의의 모형(random model)보다 예측력이 우수한지, 고려된 다른 모형들 중 어느 모형이 가장 우수한 예측력을 보유하고 있는지를 비교, 분석하는 과정이라고 할 수 있다.

제2절 모형 선택 기준 연구

모형을 선택하는 기준은 회귀모형은 Mallow's Cp, Adjusted R^2, 분류모형은 오분류율(misclassification rate)로 모형을 비교할 수 있다(김은하, 추병주, 최현수, 오미애, 김용대 등, 2015).

여러가지 모형을 이용하여 종속변수의 사후확률 $\Pr(y=1|x_1, ..., x_p)$을 계산한다. 즉, 사후확률은 독립변수 $(x_1, ..., x_p)$가 주어졌을 때, 목표변수 Y가 1이 될 확률이다. 사전확률은 독립변수를 고려하지 않은 경우의 확률 Pr(Y=1) (이 경우 자료의 분포와 모집단의 분포가 같아야 함)이고, 사후확률은 독립변수를 고려한 확률 $\Pr(y=1|x)$을 의미한다.

분류 기준값이란 사후확률을 구한 후, 이를 이용하여 자료를 분류(목표변수의 범주를 결정)하기 위하여 정하는 값이다. 예를 들어, 사후확률이 0.8 이상이면 자료를 1그룹에 할당하고 사후확률이 0.8 미만이면 자료를 0 그룹에 할당한다. 분류 기준값은 사전확률과 손실함수 등 여러 가지를 고려하여 결정하도록 한다.

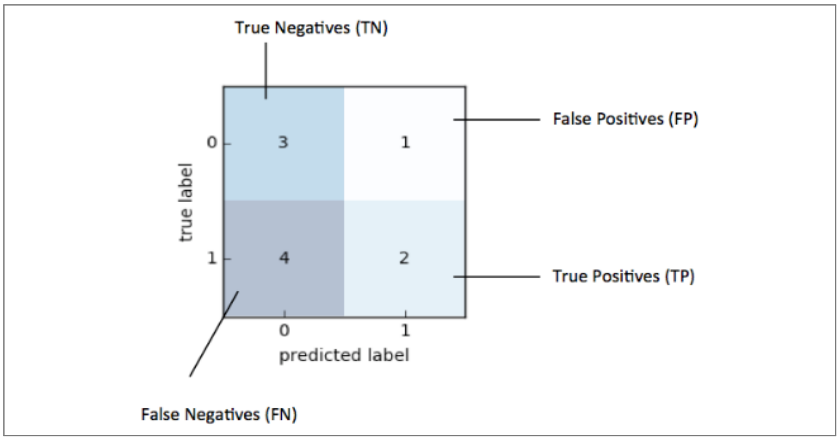
오분류표는 목표변수의 실제 범주와 모형에 의해 예측된 범주 사이의

관계를 나타내는 표이다. 오분류율=1-정분류율이며, 정분류율을 정확도(accuracy), 오분류율을 오차율(error rate)이라고 표현하기도 한다. 또한, 오분류표를 오분류 행렬(confusion matrix)이라고도 한다. 오분류율에 대한 다양한 추정치가 있을 수 있는데, 범주가 2개인 경우의 오분류표의 구성은 다음과 같다.

〈표 4-1〉 오분류표

구분		예측 변수		
		0	1	
목표변수	0	실제0, 예측0	실제0, 예측1	실제 0
	1	실제1, 예측0	실제1, 예측1	실제 1
		예측 0	예측 1	

[그림 4-1] 오분류표



자료: Raschka, S. (2017). Lift Score. https://rasbt.github.io/mlxtend/user_guide/evaluate/lift_score/에서 2017. 11. 29. 인출.

오분류율에 대한 다양한 측정치는 다음과 같다.

○ 정분류율=(실제0,예측0)빈도+(실제1,예측1)의 빈도/전체빈도

○ 오분류율=(실제0,예측1)빈도+(실제1,예측0)의 빈도/전체빈도

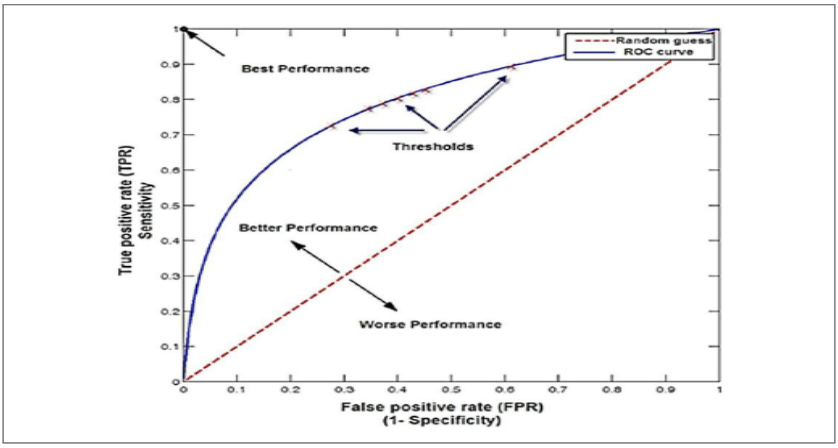
○ 민감도(sensitivity)=(실제1,예측1)의 빈도/실제 1의 빈도

○ 특이도(specificity)=(실제0,예측0)의 빈도/실제 0의 빈도

민감도는 범주 1에서의 정분류율이고, 특이도는 범주 0에서의 정분류율을 의미한다.

ROC(receiver operating characteristic) 곡선은 구축한 모형의 성능을 민감도와 특이도에 의해 판단할 수 있도록 시각화한 그림으로 신호감지이론(Signal Decision Theory)으로부터 출발하였고 신호와 잡음의 분리를 목적으로 한다.

[그림 4-2] ROC

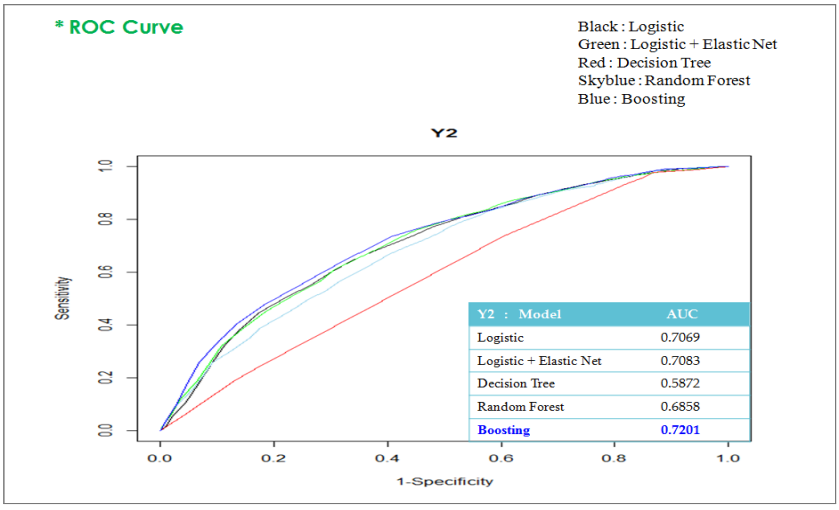


자료 : Hassouna, M., Tarhini, A., Elyas, T. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. International Business Research, Vol 8(6), pp. 224-237.

분류 기준값이 바뀌면서 특이도와 민감도의 값이 바뀌며, 여러 개의 분류 기준값에서 구하여진 민감도 특이도 값의 쌍들을 x 축에는 1-특이도, y 축에는 민감도를 지정하고 그린 곡선이다. 민감도와 특이도는 반비례하고, 따라서 ROC 곡선은 증가하는 형태를 띤다.

AUC(Area Under the Curve)는 classifier의 합리성을 나타내는 지표로 ROC 곡선의 아래 면적을 의미하며, AUC가 높을수록 좋은 모형이다.

[그림 4-3] AUC 예시



자료: Hassouna, M., Tarhini, A., Elyas, T. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. International Business Research, Vol 8(6), pp. 224-237.

이익도표(Lift Chart) 개념 역시 모형 평가 기준 중 하나이다. 이익도표 작성 과정은 다음과 같다.

우선, 모형적합을 통한 사후확률을 계산하고 사후확률의 순서에 의한 자료를 정리한다. 정렬된 자료를 균일하게 N등분하고 N등분의 각 등급에서 목표변수의 특정 범주에 대한 빈도를 구한다. N등분의 각 등급에서

%Captured Response, %Response 및 Lift 통계량을 구한다. 이익도표의 plot 은 수평축에는 N등분의 등급을, 수직축에는 위의 3개의 통계량 중 하나를 이용해 그래프를 그린다.

이익도표에 사용되는 통계량 %Captured response, %Response 수식은 다음과 같다.

%Captured response=해당 등급에서 목표변수의 특정 범주의 빈도*100/(전체에서 목표변수의 특정 범주의 빈도)

%Response=해당 등급에서 목표변수의 특정 범주의 빈도*100/(해당 등급에서 전체 빈도)

%Response 통계량은 예측확률을 순서대로 정렬하였을 때, 예측확률 상위 x% 데이터에서의 정확도를 의미하며, 일반적으로 %Response는 등급이 낮아지면서 감소한다.

Base Line Lift를 전체에서 목표변수의 특정범주의 빈도*100/(전체 빈도)로 나타낼 때, Lift는

- Lift= 해당등급의 %Response/ Base Line Lift

로 계산된다.

앞의 오분류율 표를 적용해보면 Lift는 다음과 같다.

$$Lift = \frac{(TP/(TP+FN)}{(TP+FP)/(TP+TN+FP+FN)}$$

[그림 4-4] 모형평가기준 요약

	Domain	Plot	Explanation
Lift chart	Marketing	TP Subset size	TP (TP+FP)/(TP+FP+TN+FN)
ROC curve	Communications	TP rate FP rate	TP/(TP+FN) FP/(FP+TN)
Recall-precision curve	Information retrieval	Recall Precision	TP/(TP+FN) TP/(TP+FP)

자료: Ian, H. W., & Eibe, F. (2016). Evaluation-Lift and Costs.

제 5 장

기계학습(Machine Learning) 기반 예측모형 모의분석

5 기계학습(Machine Learning) 기반 예측모형 모의분석

제1절 복지수급 예측모형 분석 DB 구축

1. 복지수급 예측모형 분석 DB 구축

여기서는 기계학습 기반 예측모형 모의분석 및 평가를 위한 분석 DB를 구축하여 가구형태, 주거 및 의료 정보, 가구의 경제적 상태, 생활여건, 가족 문제 등을 변수화하였으며 이러한 요인들이 가구의 수급 여부에 영향을 미치는지를 예측하고 이를 기반으로 예측모형을 평가하고자 하였다. 분석 DB는 2015년 말부터 보건복지부, 사회보장정보원, 한국보건사회연구원이 공동으로 연구하고, 시스템 구축을 통해 전국 지방자치단체의 일선 읍면동을 통해 시행 중인 복지사각지대 발굴관리시스템의 주요 사각지대 관련 요인 정보와 유사한 박탈 관련 문항에 대한 조사 결과들을 반영하여 구축하였다. 복지사각지대 발굴시스템은 앞서 정의한 사회보장 빅데이터를 활용하여 분석한 결과를 정책 집행에 활용하고 있는 사례이기 때문에, 모의실험을 위한 분석 DB 역시 사회보장 빅데이터와 유사하게 구성하고자 하였다. 여기서 반응변수(y)는 생계급여 수급 여부이며, 다양한 요인변수(x) 중, 먼저 가구형태는 단독가구(1인가구) 여부, 모자가구[어머니와 18세 미만의 미혼자녀로만 구성된 가구(단, 취학 중인 경우 만 22세 미만)] 여부, 부자가구[아버지와 18세 미만의 미혼자녀로만 구성된 가구(단, 취학 중인 경우 만 22세 미만)] 여부, 조손가구 또는 소년소녀가장[18세 미만의 자가 가구주로 있는 가구 또는 만 65세 이상 노인(할머니, 할아버지)과 같이 사는 경우 노인을 가구주로 하는 가구] 여부로 4가

지 유형을 고려하였다. 이와 더불어 가구원수 정보를 이용하였으며, 가구주 정보로는 가구주 교육수준, 근로능력 유무, 근로능력 정도를 활용하였다. 그리고 가구원 정보는 가구원 중에 만성질환이 있는 구성원이 있는지, 장애를 가지고 있는 구성원이 있는지를 변수화하였다. 의료위기와 관련 정보로는 건강보험료를 미납한 경험이 있는지, 연체 기간이 얼마나 되는지에 대한 정보를 활용하였다. 주거위기 관련 정보는 전세 여부, 월세 여부 등을 활용하였으며, 가구의 경제 상태를 알 수 있는 항목으로 가처분소득, 세금, 총 생활비, 총 부채, 총 재산 정보를 활용하였다. 박탈과 관련된 생활 여건으로는 세금을 못내 단전, 단수 경험이 있는지를 비롯하여 겨울에 난방을 못한 경험, 신용불량자 경험까지 다양한 환경에 대한 정보를 이용하였다. 마지막으로, 가족의 근심 걱정을 초래한 문제로 경제적 어려움, 실업 문제, 건강 문제, 알코올 문제, 주거 문제를 변수화하여 분석 DB를 구축하였다.

〈표 5-1〉 복지수급 예측모형 분석 DB 구성 항목

구분	변수설명	문항내용
반응변수	생계급여 수급 여부 (2015년 7월 1일 이후)	1.예 0.아니요
가구형태	단독가구(1인가구) 여부	1.예 / 0.아니요
가구형태	모자가구 여부	1.예 / 0.아니요
가구형태	부자가구 여부	1.예 / 0.아니요
가구형태	조손가구 또는 소년소녀가장 여부	1.예 / 0.아니요
가구정보	가구원수	명
가구주정보	가구주_교육수준	1.미취학(만7세 미만) 2.무학(만7세 이상) 3.초등학교 4.중학교 5.고등학교 6.전문대학 7.대학교 8.대학원(석사) 9.대학원(박사)

구분	변수설명	문항내용
가구주_근로능력 유무	가구주 근로능력 유무	1.예 / 0.아니요
가구주_근로능력 정도	가구주 근로능력 정도	0.근로불가 1.근로미약 2.근로가능
가구원 만성질환 유무	가구원 만성질환	0.비해당 1.만성질환 있음
가구원 장애유무	가구원 장애종류	0.비해당 1.장애인
건보미납 여부	건강보험료 미납 경험	1.있다 0.없다
건보미납기간	건강보험료 미납기간	연간: 개월
전세 여부	집의(등기상) 점유형태_전세	1.예 0.아니요
월세 여부	집의(등기상) 점유형태_월세	1.예 0.아니요
가처분소득	가처분소득	소수점 4자리(단위: 만 원), 음수는 모두 0으로
세금	세금	월평균 지출액(단위: 만 원)
총생활비	총생활비	월평균 지출액(단위: 만 원)
총부채	총부채	금융기관대출+일반사채+카드빚+전세보증금+외상및미리탄계돈+기타부채
총재산	총재산	거주주택가격+소유부동산+점유부동산+금융자산+농기계+농축산물+자동차가격+기타재산
생활여건	2달 이상 집세가 밀리거나 낼 수 없어 집을 옮긴 경험	1.있다 0.없다
	공과금을 기한 내 납부하지 못한 경험	1.있다 0.없다
	세금을 내지 못해 전기, 전화, 수도가 끊긴 경험	
	자녀 공교육비를 한 달 이상 못 준 경험	1.있다 0.없다
	돈이 없어 겨울에 난방을 못한 경험	1.있다 0.없다
	돈이 없어 본인이나 가족이 병원에 못간 경험	
	가구원 중 신용불량자가 된 경험자 여부	

구분	변수설명	문항내용
	건강보험 미납으로 인하여 보험 급여자격을 정지당한 경험 여부	1.있다 0.없다
	경제적인 어려움 때문에 먹을 것이 떨어졌는데도 더 살 돈이 없었던 경험	1.있다 0.없다
	경제적인 어려움 때문에 먹을 것을 살 돈이 없어서 균형 잡힌 식사를 할 수가 없었던 경험	1.있다 0.없다
	경제적인 어려움 때문에 먹을 것을 충분히 살 수 없어서 가구 내 성인들이 식사의 양을 줄이거나 식사를 거른 경험	1.있다 0.없다
	경제적인 어려움 때문에 먹을 것을 충분히 살 수 없어서 먹어야 한다고 생각하는 양보다 적게 먹었던 경험	1.있다 0.없다
	경제적인 어려움 때문에 먹을 것을 살 돈이 없어서 배가 고픈데도 먹지 못한 경험	1.있다 0.없다
가족경제적어려움 여부	1년간 근심이나 갈등을 초래한 문제(1순위)_경제적 어려움	1.예 0.아니요
가족취업 및 실업문제 여부	1년간 근심이나 갈등을 초래한 문제(1순위)_취업 및 실업문제	1.예 0.아니요
가족건강문제 여부	1년간 근심이나 갈등을 초래한 문제(1순위)_건강문제	1.예 0.아니요
가족알코올문제 여부	1년간 근심이나 갈등을 초래한 문제(1순위)_알코올문제	1.예 0.아니요
가족주거문제 여부	1년간 근심이나 갈등을 초래한 문제(1순위)_주거문제	1.예 0.아니요

2. 복지수급 예측모형 분석 DB의 정책적 함의

다양한 기계학습 방법론을 활용한 모형별 비교 평가를 위한 사례 분석을 위해 복지수급 예측모형 분석 DB를 구축하였다. 2015년 7월 기초생활보장제도 급여체계 개편에 따라 맞춤형 급여체계로 변화된 상황을 고려하여 기초생계급여의 수급 여부에 영향을 미치는 다양한 가구 특성 및 소득·재산 정보들을 중심으로 DB를 구축하였다.

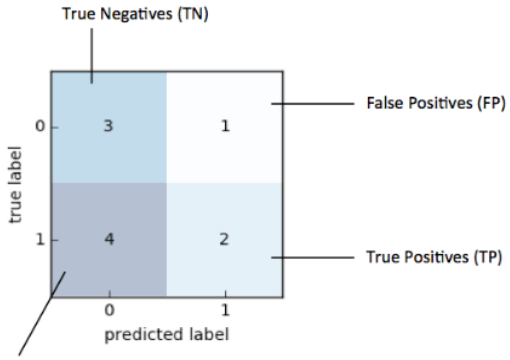
앞서, 기계학습 방법을 활용한 다양한 모형을 평가하고 선택하는 기준

으로 회귀모형에서는 Mallow's Cp, Adjusted R^2, 분류모형의 경우 오분류율(misclassification rate)과 관련된 내용을 제시하였다. 이 장에서 복지수급 예측모형 분석 DB를 활용하여 분석한 결과 역시 앞서 언급한 오분류율, 민감도, 특이도 및 이를 활용한 ROC 및 Lift chart 등을 통해서 비교 평가하고 해석할 수 있다. 그러나 사회보장 빅데이터를 기반으로 다양한 기계학습 방법을 활용하여 예측모형을 구축하고 비교 분석하는 것은 단순히 이러한 지표를 산출하고 해석하는 것에 머물러서는 안 되며 분석 결과가 관련 정책 영역에서 지닌 정책적 함의를 해석하고 도출함으로써 시사점을 제공할 수 있어야 한다. 여기서는 본 연구에서 구축한 기초생계급여 중심의 복지수급 예측모형 분석 DB를 통해 도출할 수 있는 정책적 함의에 대해 제시하고자 한다.

사례 분석을 위해 구축한 복지수급 예측모형 분석 DB를 기반으로 앞서 살펴본 다양한 기계학습 방법들을 활용하여 수급 여부에 대한 예측모형을 비교 평가하여 최적의 모형을 구축한다.

이때, 4장의 예측모형 평가방법에서 설명한 것처럼 목표 변수의 실제 범주(실제 수급 여부에 따른 구분 ☞ 수급 1, 비수급 0)와 예측모형에 의해 예측된 결과의 범주(수급 여부 예측모형에 의해 예측된 수급 여부에 따른 구분 ☞ 수급 1, 비수급 0) 사이의 관계를 나타내는 오분류표를 활용하여 다양한 지표를 산출하게 된다.

[그림 5-1] 복지수급 예측모형에 의한 오분류표의 정책적 함의



구분		예측 변수(복지수급 여부 예측결과)	
		예측 0	예측 1
목표 변수 (실제 수급 여부 정보)	실제 0	③ 실제0 & 예측0	① 실제0 & 예측1
	실제 1	④ 실제1 & 예측0	② 실제1 & 예측1

자료: Raschka, S. (2017). Lift Score. https://rasbt.github.io/mlxtend/user_guide/evaluate/lift_score/에서 2017. 11. 29. 인출.

이러한 오분류표에 제시된 바와 같이 4개 집단으로 구분되며 각각 복지수급 여부에 대한 예측결과와 관련하여 다음의 의미를 지니고 있으며 이를 통해 정책적 함의와 시사점을 도출하여 제공할 수 있다.

먼저, 집단 ②와 ③을 통해 "〈실제0,예측0〉빈도+〈실제1,예측1〉 빈도 / 전체 빈도"를 의미하는 정분류율을 산출하게 되는데, 이것은 실제 복지수급 여부 정보와 모형을 기반으로 예측된 수급 여부 예측결과가 일치하는 집단을 의미한다. 즉, 복지수급 예측모형을 통하여 수급 가능성이 높은 것으로 예측된 가구가 실제로 생계급여를 수급하고 있으며, 반대로 수급하지 못할 것으로 예측된 가구가 실제로도 수급하지 못하고 있음을 의미

한다. 이러한 정분류율이 높다면 예측모형의 정확도는 높은 것으로 평가할 수 있을 것이다. 반면에, 예측모형별 비교 평가에서 중요한 오분류율은 “1-정분류율”이며 이는 “〈실제0,예측1〉빈도+〈실제1,예측0〉빈도 / 전체 빈도”로 오차율이라고 표현할 수 있다. 무엇보다 이와 같은 오분류 사례는 민감도(〈실제1,예측1〉 빈도 / 실제 1의 빈도)와 특이도(〈실제0,예측0〉빈도 / 실제 0의 빈도) 산출에도 중요한 영향을 미치게 되지만, 복지수급 여부에 대한 예측 및 정책 집행 관련 의미는 매우 중요하다고 할 수 있다. 평가지표의 개념 정의에 따르면 민감도는 실제 범주 1의 정분류율을 특이도는 실제 범주 0에 대한 정분류율을 의미하는데, 이러한 민감도와 특이도에 해당하는 사례를 예측하는 통계적 모형에 대한 평가뿐만 아니라, 정책적으로 그렇지 않은 사례를 발견하고 이러한 사례가 어떠한 의미를 지니고 있는지 해석하여 정책적 시사점을 도출하여 제공하는 것도 중요하며 필수적이라고 할 수 있다.

제2절 기초통계 및 모형분석 결과

모형구축에 앞서 반응변수를 독립표본으로 간주하여 T-test를 실시하였다. 가구형태, 주거위기, 경제상태, 생활여건 대부분의 설명변수가 유의수준 0.05하에서 유의하였다.

〈표 5-2〉 T-test 결과

변수	Y	N	해당자 수	평균	평균의 표준오차
가구주 태어난 연도*	0	6047	.	1954.4407	.208
	1	676	.	1948.3003	.501
가구원수*	0	6047	.	2.4435	.017
	1	676	.	1.7944	.042
일인가구 여부*	0	6047	1671	.2763	.006
	1	676	359	.5311	.019
모자가구 여부*	0	6047	54	.0089	.001
	1	676	35	.0518	.009
부자가구 여부	0	6047	27	.0045	.001
	1	676	7	.0104	.004
조손가구 또는 소년소녀가장여부*	0	6047	22	.0036	.001
	1	676	17	.0251	.006
가구주 교육수준*	0	6047	.	4.6282	.022
	1	676	.	3.5917	.051
가구주 근로능력 유무*	0	6047	5438	.8993	.004
	1	676	464	.6864	.018
가구주 근로능력 정도*	0	6047	10410	1.7215	.008
	1	676	793	1.1731	.034
가구원 만성질환 유무*	0	6047	4575	.7566	.006
	1	676	624	.9231	.010
가구원 장애유무*	0	6047	1085	.1794	.005
	1	676	294	.4349	.019
건보 미납 여부*	0	6047	58	.0096	.00125
	1	676	2	.0030	.00209

변수	Y	N	해당자 수	평균	평균의 표준오차
건보 미납기간	0	6047	.	.0627	.010
	1	676	.	.0222	.018
전세 여부*	0	6047	701	.1159	.004
	1	676	63	.0932	.011
월세 여부*	0	6047	861	.1424	.004
	1	676	398	.5888	.019
가처분소득_adj*	0	6047	.	3819.0475	85.01361
	1	676	.	1488.0040	42.73412
세금(h1107_4)*	0	6047	.	12.8130	.50376
	1	676	.	.5833	.12378
총생활비(h1107_9)*	0	6047	.	295.9507	3.102
	1	676	.	120.4571	3.453
총부채	0	6047	.	21474.9883	5480.160
	1	676	.	46022.4970	25593.754
총재산*	0	5223	.	39396.1324	3923.039
	1	540	.	5273.5519	1876.486
2달 이상 집세가 밀리거나 낼 수 없어 집을 옮긴 경험*	0	6047	25	.0041	.001
	1	676	19	.0281	.006
공과금을 기한 내 납부하지 못한 경험*	0	6047	115	.0190	.002
	1	676	25	.0370	.007
세금을 내지 못해 전기, 전화, 수도가 끊긴 경험	0	6047	16	.0026	.001
	1	676	4	.0059	.003
자녀의 공교육비를 한 달 이상 못 준 경험	0	6047	14	.0023	.001
	1	676	1	.0015	.001
돈이 없어 겨울에 난방을 못한 경험*	0	6047	68	.0112	.001
	1	676	35	.0518	.009
돈이 없어 본인이나 가족이 병원에 못 간 경험*	0	6047	53	.0088	.001
	1	676	22	.0325	.007
가구원 중 신용불량자가 된 경험자 여부*	0	6047	155	.0256	.002
	1	676	63	.0932	.011
건강보험 미납으로 인하여 보험 급여자격을 정지당한 경험 여부	0	6047	38	.0063	.001
	1	676	2	.0030	.002

변수	Y	N	해당자 수	평균	평균의 표준오차
경제적인 어려움 때문에 먹을 것이 떨어졌는데도 더 살 돈이 없었던 경험*	0	6047	138	.0228	.002
	1	676	89	.1317	.013
경제적인 어려움 때문에 먹을 것을 살 돈이 없어서 균형 잡힌 식사를 할 수가 없었던 경험*	0	6047	291	.0481	.003
	1	676	191	.2825	.017
경제적인 어려움 때문에 먹을 것을 충분히 살 수 없어서 가구 내 성인들이 식사의 양을 줄이거나 식사를 거른 경험*	0	6047	25	.0041	.001
	1	676	18	.0266	.006
경제적인 어려움 때문에 먹을 것을 충분히 살 수 없어서 먹어야 한다고 생각하는 양보다 적게 먹었던 경험*	0	6047	39	.0064	.001
	1	676	31	.0459	.008
경제적인 어려움 때문에 먹을 것을 살 돈이 없어서 배가 고픈데도 먹지 못한 경험*	0	6047	21	.0035	.001
	1	676	14	.0207	.005
가족 경제적어려움 여부*	0	6047	680	.1125	.004
	1	676	179	.2648	.017
가족 취업 및 실업문제 여부	0	6047	205	.0339	.002
	1	676	18	.0266	.006
가족 건강문제 여부*	0	6047	1733	.2866	.006
	1	676	284	.4201	.019
가족 알코올문제 여부	0	6047	28	.0046	.001
	1	676	3	.0044	.003
가족 주거문제 여부	0	6047	45	.0074	.001
	1	676	10	.0148	.005

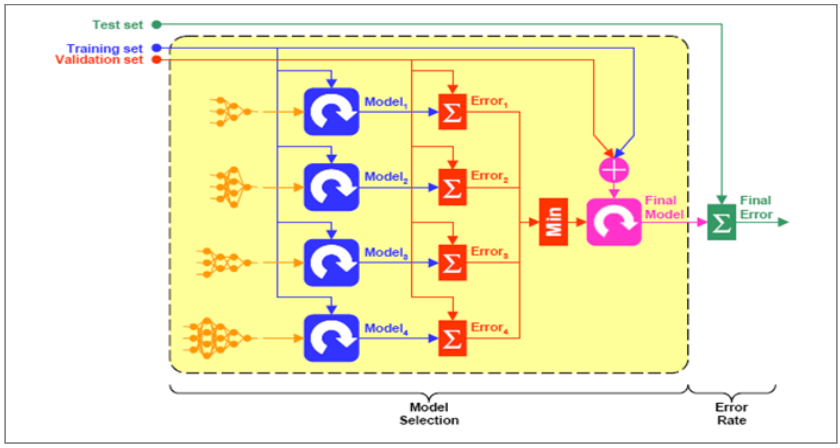
주: 유의수준 5%하에서 유의하면 변수에 * 표시를 하였음.

분석 모형은 로지스틱 회귀모형, Elastic Net, Tree, Boosting, Random Forest, SVM, Deep Learning이다. 모형 구축을 위한 분석에서 총재산 항목은 결측치가 많이 존재하여 분석변수에서는 제외하였다.

그리고 t-test에서 유의하지 않은 변수들도 모형 간의 비교 분석을 위해 분석 변수에 포함하여 모형을 구축하였다.

데이터는 모형을 구축하는 데이터와 모형을 선택하는 데이터, 모형을 평가하는 데이터 3가지로 나누어진다. 때로는 모형을 선택하는 데이터는 모형을 구축하는 데이터와 합쳐서 모형구축 데이터(Training data), 모형평가 데이터(Test data)로 나누기도 한다.

[그림 5-2] 모형구축과 모형평가 데이터



여기에서는 10 fold cross-validation을 이용하여 모형을 구축하고 모형을 평가하였다. 전체 데이터셋을 10개로 나누고, 첫 실험에는 첫 번째 데이터셋을 모형 테스트에 사용하고 나머지 9개 데이터셋을 모형 구축에 사용한다. 두 번째 실험은 두 번째 데이터셋을 모형 평가에 사용하고 나머지 9개 데이터셋을 모형 구축에 사용한다. 이렇게 10번의 실험을 다 거치면 모든 데이터는 모형 구축에 1번씩만 사용하게 되고, 모형 구축 데이터셋과 모형 평가 데이터셋이 겹치는 일이 없게 된다.

〈표 5-3〉 10 fold Cross-Validation

Experiment1	T1									
Experiment2		T2								
Experiment3			T3							
Experiment4				T4						
Experiment5					T5					
Experiment6						T6				
Experiment7							T7			
Experiment8								T8		
Experiment9									T9	
Experiment10										T10

구축 모형은 각 모형마다 10개의 모형이 구축된다.

실제 모형 평가에서는 10fold CV로 모형 평가 데이터를 구축하였다.

아래 모형의 계수 또는 그림은 각 모형의 특성을 살펴보기 위해, 전체 데이터로 모형을 구축하여 제시하였다.

Logistic 회귀계수를 살펴보면 다음과 같다.

〈표 5-4〉 Logistic Coeff

변수	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-3.7805	0.3552	0
가구원수	0.8233	0.0976	0
일인가구 여부	0.7584	0.1681	0
모자가구 여부	2.1959	0.3046	0
부자가구 여부	1.022	0.5642	0.0701
조손가구 또는 소년소녀가장 여부	2.4304	0.4225	0
가구주_교육수준	-0.0748	0.0436	0.0864
가구주 근로 능력 유무	0.4694	0.2718	0.0842
가구주 근로 능력 정도	-0.5855	0.1456	1.00E-04
가구원 만성질환	0.8132	0.1884	0
가구원 장애종류	1.0777	0.1107	0
건강보험료 미납경험	-2.8878	1.5655	0.0651
건강보험료 미납기간	0.0805	0.192	0.6751
집의(등기상) 점유형태_전세	1.0055	0.1738	0

변수	Estimate	Std. Error	Pr(>|z|)
집의(등기상) 점유형태_월세	2.2412	0.1209	0
가처분소득	-1.00E-04	1.00E-04	0.1857
세금	-0.1173	0.0341	6.00E-04
총생활비	-0.0081	0.0015	0
총부채	0	0	0.4999
2달 이상 집세가 밀리거나 낼 수 없어 집을 옮긴 경험	-0.1311	0.4079	0.748
공과금을 기한 내 납부하지 못한 경험	-0.3631	0.3598	0.3129
세금을 내지 못해 전기, 전화, 수도가 끊긴 경험	0.4165	0.7931	0.5995
자녀의 공교육비를 한 달 이상 못 준 경험	-0.6708	1.1427	0.5572
돈이 없어 겨울에 난방을 못 한 경험	-0.0862	0.293	0.7686
돈이 없어 본인이나 가족이 병원에 못 간 경험	-0.7983	0.3325	0.0164
가구원 중 신용불량자가 된 경험자 여부	0.4696	0.2206	0.0333
건강보험 미납으로 인하여 보험 급여자격을 정지당한 경험 여부	0.6354	1.2976	0.6244
경제적인 어려움 때문에 먹을 것이 떨어졌는데도 더 살 돈이 없었던 경험	0.1754	0.2372	0.4596
경제적인 어려움 때문에 먹을 것을 살 돈이 없어서 균형 잡힌 식사를 할 수가 없었던 경험	0.6464	0.1682	1.00E-04
경제적인 어려움 때문에 먹을 것을 충분히 살 수 없어서 가구 내 성인들이 식사의 양을 줄이거나 식사를 거른 경험	-0.049	0.5213	0.9252
경제적인 어려움 때문에 먹을 것을 충분히 살 수 없어서 먹어야 한다고 생각하는 양보다 적게 먹었던 경험	0.2915	0.4143	0.4816
경제적인 어려움 때문에 먹을 것을 살 돈이 없어서 배가 고픈데도 먹지 못한 경험	-0.3686	0.5761	0.5223
1년간 근심이나 갈등을 초래한 문제(1순위)_경제적 어려움	0.324	0.1508	0.0317
1년간 근심이나 갈등을 초래한 문제(1순위)_취업 및 실업문제	0.2843	0.313	0.3636
1년간 근심이나 갈등을 초래한 문제(1순위)_건강문제	0.1868	0.1234	0.1303
1년간 근심이나 갈등을 초래한 문제(1순위)_알코올문제	0.4284	0.6973	0.539
1년간 근심이나 갈등을 초래한 문제(1순위)_주거문제	0.8481	0.4629	0.067

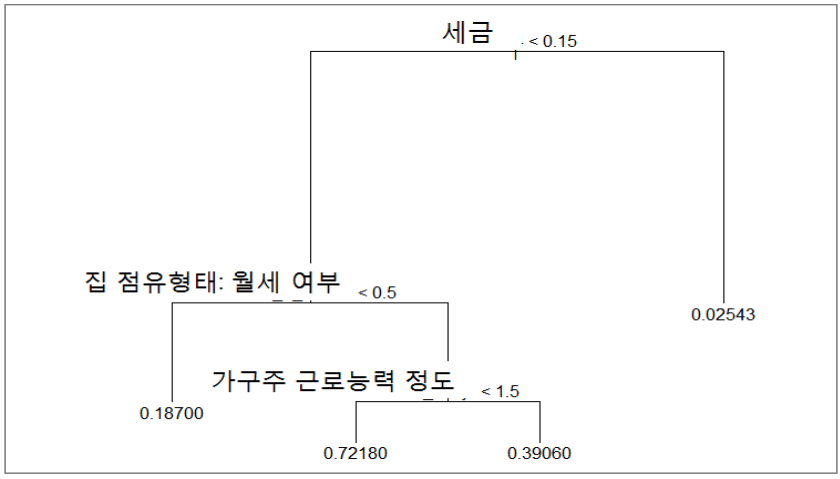
Elastic Net 의 회귀계수를 살펴보면 Logitic 회귀계수와 크게 다르지 않음을 확인할 수 있다.

〈표 5-5〉 Elastic Net Coeff

Elastic Net	Estimate
beta0	-3.7457
가구원수	0.8118
일인가구 여부	0.746
모자가구 여부	2.2011
부자가구 여부	1.0226
조손가구 또는 소년소녀가장 여부	2.4354
가구주_교육수준	-0.0779
가구주 근로 능력 유무	0.4584
가구주 근로 능력 정도	-0.5822
가구원 만성질환	0.8099
가구원 장애종류	1.0748
건강보험료 미납경험	-2.7581
건강보험료 미납기간	0.0673
집의(등기상) 점유형태_전세	1.0116
집의(등기상) 점유형태_월세	2.249
가처분소득	-1.00E-04
세금	-0.0982
총생활비	-0.0082
총부채	0
2달 이상 집세가 밀리거나 낼 수 없어 집을 옮긴 경험	-0.1293
공과금을 기한 내 납부하지 못 한 경험	-0.3548
세금을 내지 못해 전기, 전화, 수도가 끊긴 경험	0.3948
자녀의 공교육비를 한 달 이상 못 준 경험	-0.6282
돈이 없어 겨울에 난방을 못 한 경험	-0.0843
돈이 없어 본인이나 가족이 병원에 못 간 경험	-0.7912
가구원 중 신용불량자가 된 경험자 여부	0.4716
건강보험 미납으로 인하여 보험 급여자격을 정지당한 경험 여부	0.6064
경제적인 어려움 때문에 먹을 것이 떨어졌는데도 더 살 돈이 없었던 경험	0.1728

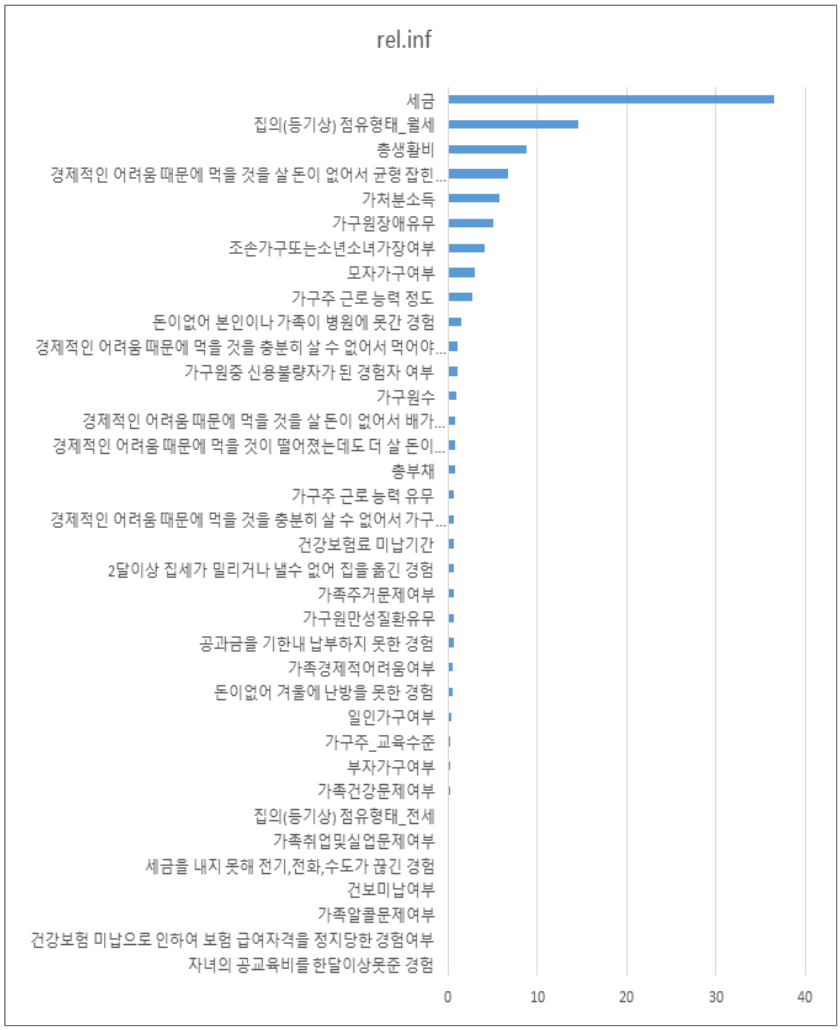
Elastic Net	Estimate
경제적인 어려움 때문에 먹을 것을 살 돈이 없어서 균형 잡힌 식사를 할 수가 없었던 경험	0.6501
경제적인 어려움 때문에 먹을 것을 충분히 살 수 없어서 가구 내 성인들이 식사의 양을 줄이거나 식사를 거른 경험	-0.0464
경제적인 어려움 때문에 먹을 것을 충분히 살 수 없어서 먹어야 한다고 생각하는 양보다 적게 먹었던 경험	0.2841
경제적인 어려움 때문에 먹을 것을 살 돈이 없어서 배가 고픈데도 먹지 못한 경험	-0.3584
1년간 근심이나 갈등을 초래한 문제(1순위)_경제적 어려움	0.3241
1년간 근심이나 갈등을 초래한 문제(1순위)_취업 및 실업문제	0.2822
1년간 근심이나 갈등을 초래한 문제(1순위)_건강문제	0.1846
1년간 근심이나 갈등을 초래한 문제(1순위)_알코올문제	0.4172
1년간 근심이나 갈등을 초래한 문제(1순위)_주거문제	0.8399

[그림 5-3] Decision Tree 모형 주요 변수



의사결정나무 모형에서 수급가구를 예측하는 데 주요한 변수는 세금을 얼마 내는지, 집 점유형태가 월세인지, 가구주 근로능력 수준은 어느 정도인지이다.

[그림 5-4] Boosting 모형(상대적 영향도)



부스팅 알고리즘에서 수급 여부를 결정하는 데 있어서 중요한 변수는 세금, 월세 여부, 총생활비, 경제적 어려움으로 균형잡힌 식사 못한 경험, 가처분소득, 가구원장애 유무, 가구형태 정보, 가구주 근로능력 정보이다.

제3절 기계학습 기반 예측모형 비교 및 평가

1. 오분류표

분류기준을 0.5로 했을 때 각 모형별 오분류표는 다음과 같다.

〈표 5-6〉 로지스틱(분류기준: 0.5) 오분류표

구분		예측 변수		
		0	1	
목표변수	0	5,916	131	실제 0
	1	413	263	실제 1
		예측 0	예측 1	

〈표 5-7〉 Elastic Net(분류기준: 0.5) 오분류표

구분		예측 변수		
		0	1	
목표변수	0	5,918	129	실제 0
	1	415	261	실제 1
		예측 0	예측 1	

〈표 5-8〉 Tree(분류기준: 0.5) 오분류표

구분		예측 변수		
		0	1	
목표변수	0	5,968	79	실제 0
	1	471	205	실제 1
		예측 0	예측 1	

〈표 5-9〉 Boosting(분류기준: 0.5) 오분류표

구분		예측 변수		
		0	1	
목표변수	0	5,911	136	실제 0
	1	376	300	실제 1
		예측 0	예측 1	

〈표 5-10〉 Random Forest(분류기준: 0.5) 오분류표

구분		예측 변수		
		0	1	
목표변수	0	5,893	154	실제 0
	1	389	287	실제 1
		예측 0	예측 1	

〈표 5-11〉 SVM(분류기준: 0.5) 오분류표

구분		예측 변수		
		0	1	
목표변수	0	5,944	103	실제 0
	1	486	190	실제 1
		예측 0	예측 1	

〈표 5-12〉 Deep Learning(CNN) (분류기준: 0.5) 오분류표

구분		예측 변수		
		0	1	
목표변수	0	5,795	252	실제 0
	1	367	309	실제 1
		예측 0	예측 1	

분류기준을 0.5로 했을 때, 정확도는 부스팅이 92.38%로 가장 높았다.

2. 정확도, 특이도 및 민감도

분류문제에서 분류기준을 얼마로 잡느냐에 따라 정확도, 특이도, 민감도는 달라진다. 분류기준을 0.5, 0.6, 0.7, 0.8, 0.9로 했을 때의 모형별 정확도, 특이도, 민감도 결과는 다음과 같다.

〈표 5-13〉 분류기준 0.5 결과

	logistic	elastic	tree	boosting	random forest	svm	deep learning
정확도	91.91%	91.91%	91.82%	92.38%	91.92%	91.24%	90.79%
특이도	97.83%	97.87%	98.69%	97.75%	97.45%	98.29%	95.83%
민감도	38.91%	38.61%	30.33%	44.38%	42.46%	28.10%	45.71%

〈표 5-14〉 분류기준 0.6 결과

	logistic	elastic	tree	boosting	random forest	svm	deep learning
정확도	91.79%	91.80%	91.82%	92.09%	92.16%	91.16%	91.00%
특이도	98.86%	98.86%	98.69%	98.68%	98.66%	98.71%	96.79%
민감도	28.55%	28.70%	30.33%	33.14%	34.02%	23.66%	39.20%

〈표 5-15〉 분류기준 0.7 결과

	logistic	elastic	tree	boosting	random forest	svm	deep learning
정확도	91.21%	91.19%	91.82%	91.39%	91.54%	90.88%	90.94%
특이도	99.29%	99.27%	98.69%	99.22%	99.22%	98.90%	97.53%
민감도	18.93%	18.93%	30.33%	21.30%	22.78%	19.08%	31.95%

평가지표를 민감도로 한다면 이 경우에는 Decision Tree의 결과가 가장 좋다. 하지만 Decision Tree의 경우, 분류기준을 0.5부터 0.7로 했을

경우, 실제 1인데 1로 예측되는 케이스 숫자 205명이다. 반면, 분류기준을 0.8, 0.9로 했을 경우에는 Decision Tree 모형은 모두 0으로 분류(예측)하였다.

이를 보면 반응변수가 0의 케이스 숫자와 1의 케이스 숫자가 많이 차이가 나는 경우, 하나의 평가 기준만 고려해서는 안 된다는 것을 알 수 있다.

〈표 5-16〉 분류기준 0.8 결과

	logistic	elastic	tree	boosting	random forest	svm	deep learning
정확도	90.41%	90.41%	89.94%	90.79%	90.79%	90.82%	90.86%
특이도	99.69%	99.69%	100%	99.69%	99.77%	99.47%	98.31%
민감도	7.40%	7.40%	0%	11.24%	10.50%	13.46%	24.26%

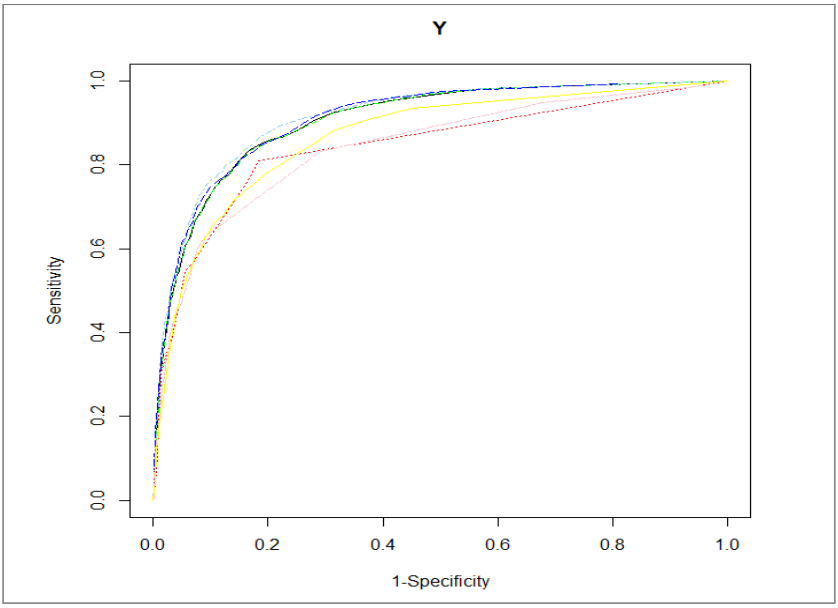
〈표 5-17〉 분류기준 0.9 결과

	logistic	elastic	tree	boosting	random forest	svm	deep learning
정확도	90.06%	90.05%	89.94%	90.14%	90.08%	90.45%	90.74%
특이도	99.95%	99.95%	100%	99.93%	99.95%	99.76%	99.02%
민감도	1.63%	1.48%	0%	2.51%	1.78%	7.10%	16.71%

분류기준에 따라 정확도, 특이도, 민감도 결과 부스팅이 전반적으로 우수한 성능을 보이고 있다. deep learning(CNN)의 경우, 민감도 결과를 보면 다른 방법에 비해 상대적으로 높은 성능을 보이고 있음을 확인하였다.

3. ROC Curve

[그림 5-5] ROC Curve



4. AUC

〈표 5-18〉 AUC 결과

모형	AUC
Logistic	0.9058
Elastic Net	0.9059
Decision Tree	0.8335
Boodsting	0.9161
Random Forest	0.9101
SVM	0.8530
Deep Learning(CNN)	0.8726

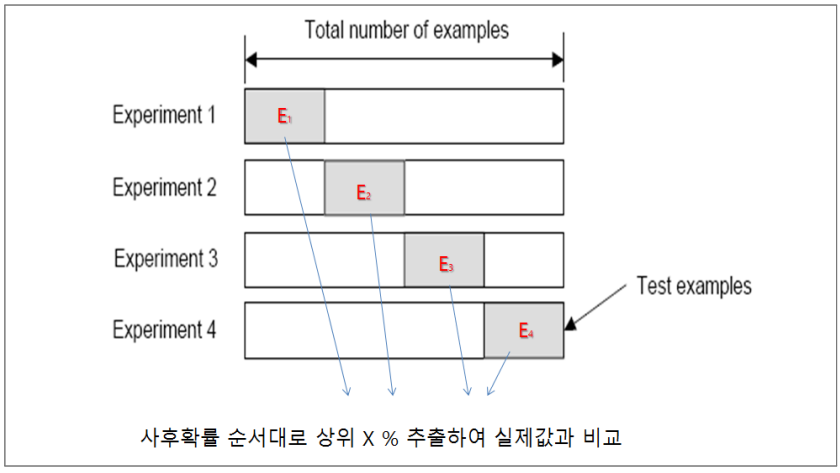
ROC, AUC 모두 앙상블 기법인 부스팅과 랜덤 포레스트의 결과가 제일 좋음을 알 수 있다.

5. Lift

이익도표의 %Response 통계량으로 모형을 평가하였다.

사후확률 순서대로 정렬하였을 때 상위 5%, 상위 10%, 상위 15%, 상위 20%, 상위 25%, 상위 30%를 1등급으로 하여 모형별로 %Response 통계량을 계산하였다.

〔그림 5-6〕 Lift 산출 과정



10 fold partition으로 10개 test set의 %Response 평균 산출한 결과는 다음과 같다.

〈표 5-19〉 등급별 lift의 %Response

등급	logistic	elastic	Tree	Boosting	Ranf	svm	Deep Learning	random
5%	11.7	23.6	24.1	24.6	24.3	12.9	14.5	3.4
10%	29.5	41.1	44.4	44.4	41	32.9	32.6	8
15%	55	66.9	70.4	68	65.7	57.3	56.8	10.3
20%	81.6	94.4	98.1	96.2	94.2	84.4	82.4	14.3
25%	111.4	124.3	126.2	126.3	124.5	113.4	110.7	16.8
30%	141.6	156.1	159.4	157.6	156.2	144.2	138.9	20.7

〈표 5-20〉 등급별 lift의 %Response 정확도(%)

	logistic	elastic	Tree	Boosting	Ranf	svm	Deep Learning	random
5% acc	34.41	69.41	70.88	72.35	71.47	37.94	42.65	3.4
10% acc	44.03	61.34	66.27	66.27	61.19	49.10	48.66	8
15% acc	54.46	66.24	69.7	67.33	65.05	56.73	56.24	10.3
20% acc	60.9	70.45	73.21	71.79	70.3	62.99	61.49	14.3
25% acc	66.31	73.99	75.12	75.18	74.11	67.50	65.89	16.8
30% acc	70.1	77.28	78.91	78.02	77.33	71.39	68.76	20.7

lift 의 %Response 통계량 역시, 부스팅과 랜덤 포레스트 알고리즘이 높은 성능을 보이고 있고, 의사결정나무도 상대적으로 좋은 결과를 보이고 있다. 이 데이터에서는 Deep learning(CNN) 방법이 이미지(image) 분석이나 언어(language) 분석처럼 좋은 결과를 나타내지는 못했다.

제4절 소결

이 장에서는 사회보장 빅데이터와 유사한 형태의 분석 DB를 구축하여 가구의 생계급여 수급 여부에 영향을 미치는 요인들을 분석하고, 여러 예측모형을 비교·분석하여 다양한 모형평가 기준으로 결과를 제시하였다. 예측모형을 평가하기 위하여 10fold CV를 이용하였기 때문에 예측모형의 과적합 문제는 발생하지 않는다.

오분류율, AUC, ROC, Lift의 %Response 모두 기계학습 기법인 부스팅의 결과가 제일 좋음을 확인할 수 있었다. 최신 기법인 딥러닝(CNN)의 경우, 다른 분석 결과에 비해 민감도가 높은 편이다. 이는 복지 수급이 필요한 사람을 다소 과대해서 예측한 결과에서 비롯된 것으로, CNN 결과를 바탕으로 복지 수급 여부를 결정하면 복지 수급이 필요한 사람이 복지 수급을 받지 못할 확률은 줄어들 수 있다. 하지만 반대로 복지 수급을 받지 않아도 되는 사람까지 복지 서비스를 받을 수 있기 때문에 정부 입장에서는 재정적 부담이 커진다. 따라서, 분석 모델의 선택은 단순한 정확도에 의한 판단 이외에도 중요한 요인이 무엇인가에 따라 달라질 수 있다. 딥러닝 기법은 종속변수가 범주형(binary)인 데이터인 분석DB에서는 다른 기계학습 기법에 비해 좋은 성능을 보이지는 못했다. 분석 DB에는 독립변수에도 범주형 변수가 많이 존재하기 때문에 이런 데이터에도 딥러닝의 예측결과가 좋을지에 대한 부분은 더 많은 연구가 필요하다.

Lift의 %Response(test 데이터의 확률값 상위 5%)로 종속변수가 범주형일 때 적용할 수 있는 logistic 회귀모형과 부스팅 결과를 비교하면 logistic 방법을 적용했을 때보다 부스팅 방법의 성능이 2배 이상 되는 것을 확인할 수 있다.

복지사각지대 발굴시스템 사례를 예로 들면, 모형에서 정확도에서 1%

차이가 난다는 것은 1만 명의 위기가구 대상자 리스트를 지자체에 제공하였을 때, 정확도가 1% 높은 모형을 적용했을 경우 100명을 더 발굴할 수 있다는 것을 의미한다. 정확도 1% 의 차이는 모형 평가 부분에서는 크게 차이가 나지 않는다고 할 수도 있지만, 행정 집행과 관련된 효율 측면에서는 큰 차이를 보일 수 있다.

복지수급 예측 관련 분석 DB를 분석함으로써 도출할 수 있는 정책적 함의와 시사점은 다음과 같다. 복지사각지대의 예측을 통한 복지대상자의 발굴 및 확인뿐만 아니라, 수급 여부 예측 결과를 바탕으로 실제 수급 여부와의 비교를 통하여 기초생활보장제도 개별 급여뿐만 아니라 각종 차상위 지원제도, 기초연금 및 장애인연금 등 다양한 복지제도의 수급률에 대한 통계적 추정을 가능하게 함으로써 실증적 근거를 기반으로 제도 개선을 통한 수급률 제고방안 마련 등 정책적 측면에서 실질적으로 활용할 수 있다는 점이다. 이처럼 사회보장 빅데이터를 기반으로 구축된 다양한 예측모형은 정책 영역에서 중요한 의미와 함께 활용 가능하다고 할 수 있다.

제 6 장 결론 및 정책 제언

6 결론 및 정책 제언

제1절 기계학습 기법 관련 이슈

현재 우리는 자료가 넘쳐나는 시대에 살고 있다. 정보통신기술(ICT)의 발달로 모든 방면(사람, 기계, 장소, 시간)에서 자료가 쌓이고 소비되고 있으며 이렇게 수집·사용된 방대한 자료들을 통해 유용한 정보를 추출하여 교류하고 상호 작용하는 지능형 인프라 및 서비스 기술 개발이 중요한 과제로 떠오르고 있다. 서로 다른 매체를 통해 수집된 다양한 형태의 자료들은 혁신적이고 삶의 질을 향상시킬 수 있는 방향으로 합성(fused)되고 탐색되어야 한다. 이 때문에 이 과정에서 자료분석 전문가들(Data Scientists)의 엄청난 노력과 시간이 투자되고 자료가 복잡할수록 이러한 절차는 더욱 전문가의 판단이나 경험 등 자동화할 수 없는 요소가 지배적이게 된다. 그러나 인간의 통찰력으로는 이러한 방대하고 복잡 다양한 자료들이 만들어지고 축적되는 속도를 따라 정보를 추출해 내기에 역부족이며 부족한 전문가 인력 수로 인해 진화하는 기술에 상응할 자기학습적 모형(self-learning model) 혹은 개개인의 기호에 적합(personalized/adaptive)한 서비스 구축에 필요한 규모의 경제(scalability)를 얻기 어렵다. 특히 자료분석 업무(Data Science Project) 중 80%에 달하는 시간은 대부분 자료 분석 전문가의 수작업이 필요한 자료의 전처리(pre-processing), 탐색(exploration), 그리고 특성추출(feature engineering)에 소비된다. 현재 상용되는 중앙집중적 빅데이터 기반(architectures)은 모든 것을 축적하는 동시에 그만큼 잊혀지는 자료를

발생시키며 자료를 전달하고 보고하는 기능에는 최적화 되었지만 자료 학습 능력에는 시간차에 의한 정보 전달 지연 현상이 누적될 수밖에 없으며 개인화된 서비스 수요를 감당할 자료에 기반을 둔 의사결정기술과는 맞지 않는다는 문제가 있다. 이를 극복하고자 하는 노력으로 현재 주목받고 있는 기술로는 딥러닝(deep learning)이 있다. 이는 사물이나 자료를 군집화하거나 분류하는 데 사용하는 기술이다. 딥러닝의 핵심은 분류를 통한 예측이다. 인간의 두뇌가 수많은 자료 속에서 패턴을 발견한 뒤 사물을 구분하는 정보처리 방식을 모방해 컴퓨터가 사물을 분별하도록 기계를 학습시킨다. 그러나 많은 자료들은 학습의 기반이 될 분류 레이블(label)이 없거나 있다고 해도 제한적이어서 딥러닝의 방대한 학습에 필수로 요구되는 분류 표본(sample)이 부족하기 때문에 오분류의 확률이 높다. 이 때문에 적은 표본의 분류된 자료나 전혀 분류 레이블의 습득이 불가능한 경우의 자료에서도 잘 작동할 수 있는 방법의 개발이 필수적이다. 자동화된 기계학습(auto ML: automated machine learning)은 자료를 효율적으로 사용할 수 있고 비지도 학습(unsupervised learning)에 초점을 맞춘 접근법이다. 만약 완전히 자동화된 기계학습이 가능하다면 자료분석 과정에서 가장 심각한 병목현상(bottleneck)으로 꼽히는 전처리/탐색/특성추출 절차에 소요되는 전문가의 시간과 노력을 획기적으로 줄일 수 있으며 어떠한 복잡 다양한 자료도 인간이 인식하는 의미 있는 형태(human understandable context)로 쉽게 변환될 수 있는 기회가 열리게 된다.

이렇듯, 기계학습 기술은 빠르게 발전하고 있지만, 기계학습의 발전이 사회 전반에 걸쳐 어떠한 영향을 미치게 될지에 대한 연구는 전 세계적으로 진행 중에 있다. 기계학습이 가져다 줄 수 있는 장점을 극대화하기 위해서는 다음의 사항(고학수, 2017, pp. 77-78)들을 고려해 보아야 한다.

이러한 이슈들은 기계학습 기술이 발전하고 정책 영역에서도 광범위하게 적용되기 위해서 검토가 선행되어야 하는 필수적인 요소라고 할 수 있다.

우선, 개방형 연구 관행(Open research norms and practices)이 필요하다. 기계학습이 빠르게 발전한 이유 중 하나는 연구 결과를 공개하고 공유하고 장려하고 있기 때문이라고 할 수 있다. 기계학습 기법을 적극적으로 활용하고 있는 구글의 경우, 전 세계 연구 커뮤니티에 참여하여 여러 연구 결과를 발표하고, 다양한 주제에 대한 콘퍼런스에 참여하고 있다. 또한, 구글 내부적으로 사용해 온 기계학습 툴킷인 텐서플로를 개방하여 전 세계의 연구자들과 개발자들이 데이터 분석을 하고 최첨단 기술을 발전시키는 것에 기여하고 있다. 딥마인드 역시 연구의 활성화를 위해 알파고(AlphaGo)의 대국을 전문가의 해설과 함께 게시하고, 여러 언어로 이를 제공하였다. 개방형 연구에 대한 공감대가 확산된다면 기계학습은 지금보다 더 빠르게 발전할 수 있을 것이다.

다음으로, 기계학습이 안전하고 윤리적인 방향으로 발전할 수 있도록 해야 한다. 모든 기술이 그렇듯이, 기계학습의 경우에도 이점을 극대화하고 부작용을 최소화하는 것이 중요하다. 기계학습 분야는 최근에 급속도로 발전하였기 때문에 다양한 상황에 기계학습이 적용되는 경우 어떤 위험이 발생할지는 아직 아무도 예측할 수 없다. 편향된 데이터에 기반을 둔 기계학습은 편향되거나 잘못된 결과를 도출할 수 있는 반면, 객관적인 기준에 맞춰진 기계학습은 오히려 이런 편향을 줄일 수 있다. 완벽한 시스템은 존재하지 않고, 어떤 시스템 안에서도 오류는 존재하지만 기술의 발전으로 이런 문제들이 해결될 수 있도록 노력해야 한다. 이를 위해서 연구자들은 문제를 해결하는 데 도움이 될 수 있는 통계학적 지식과 인문학적 사고를 갖추어야 한다. 이미 해외에서는 알고리즘의 편향성과 이를 수정하는 방법에 대한 연구를 진행하고 있다. 그리고 윤리적이고 책임 있

는 알고리즘 개발에 대한 연구자 대상의 교육과정도 필요하다.

마지막으로, 기계학습이 정책 영역을 포함한 다양한 분야에 접목되어 활용될 수 있도록 해야 한다. 기계학습은 사회 전체적으로 널리 공유될 수 있는 이익을 창출할 수 있다. 기계학습의 잠재적 문제들은 기술 개발에 대한 다양한 관점, 배경, 경험을 가진 사람들과의 협업을 통해 더욱 발전할 수 있다. 따라서 통계학과 컴퓨터공학 교육에 대한 지속적 투자가 필요하며, 기계학습을 배우고자 하는 사람들에게 노하우를 전수하고 학습 자료를 제공하는 것 역시 중요하다. 인공지능과 함께 새로운 문제들은 지속적으로 발생할 수밖에 없기 때문에 학계, 정책 입안자, 민간과 공공부문 전문가, 산업 분야의 이해관계자들의 기계학습에 대한 관심과 참여가 필요하다.

제2절 정책 제언

이 연구에서는 앞서 정의한 사회보장 빅데이터와 유사한 분석DB를 구축하여 딥러닝을 포함한 다양한 기계학습 통계기법을 적용하고 여러 모형 선택기준으로 최적의 예측모형을 살펴보았다는 점에서 기존 연구와 차별성이 있다. 분석DB에서는 기계학습 기법인 부스팅의 결과가 제일 좋은 성능을 보여주었다. 딥러닝은 이미지 및 언어 분석에서는 탁월한 성능을 보여주지만, 분석DB에서는 부스팅과 랜덤 포레스트의 결과보다 좋지 않음을 알 수 있다. 이는 모든 데이터에서 하나의 기계학습 기법이 좋은 성능을 보이는 것은 아니라는 것을 의미한다. 부스팅 기법은 대부분의 경우 좋은 예측력을 보여주지만, 분석 데이터의 속성에 따라서 다른 기계학습 방법이 더 좋을 수 있다. 사회보장 빅데이터가 이미지 및 영상 데이터인 경우에는 부스팅 방법보다는 딥러닝 방법이 더 적합할 수 있다. 따라

서 기계학습 기법을 적용하려면 주어진 문제를 해결하기 위해 구축된 사회보장 빅데이터의 속성을 우선 이해해야 할 것이다. 복지사각지대 발굴 시스템의 경우 복지수급 대상자와 복지수급 비대상자의 인구학적 특성, 경제적 특성, 주거 특성 등에 대한 정보가 필요하다. 위기아동 발굴 시스템은 학대 아동과 일반 아동의 부모 특성, 아동 특성이 중요한 정보일 것이다. 보건의료 분야의 영상데이터의 경우에는 이상자료(abnormal data)의 특성을 파악할 수 있어야 한다.

기계학습이 보건복지 정책 분야에서 긍정적인 영향을 미칠 수 있는지를 파악하기 위해서는 먼저 기술 전반에 대한 이해가 선행되어야 한다. 정부는 기계학습과 관련된 이슈들을 해결하기 위한 다양한 연구를 지원하고 기술을 개발하고 적용하는 과정에 중요한 역할을 할 수 있다. 정부는 사회적인 문제들을 해결하기 위해 어떤 방식으로 기계학습 방법을 적용해야 하는지, 어떻게 기술의 잠재적인 한계를 극복할 수 있는지에 대한 연구를 지원해야 한다.

또한, 기계학습이 에너지, 운송, 환경, 복지, 공중보건 등의 다양한 분야에서 전 세계적인 문제를 해결할 수 있는 효과적인 도구라는 것은 이미 입증된 부분이다. 따라서 공공과 민간부문 간의 협업과 다수의 이해관계자들의 참여를 유도해야 한다. 국가적 차원에서 여러 분야를 포괄하는 태스크포스를 만들고, 산업계를 포함한 다수의 이해관계자들이 모여서 이슈 및 모범 사례를 공유할 수 있어야 한다.

한편, 기계학습 분야는 전 세계 연구자들의 참여와 광범위한 데이터의 활용이 이루어지는 국제적인 분야로, 기계학습의 성공은 정보, 통신, 기술 정책에서부터 디지털경제, 저작권에 이르기까지 상호 연결된 분야에서 얼마나 많은 혁신 친화적인 법률 제도들이 만들어지는지에 달려있다. 다수의 이해관계자들의 참여는 국내의 법 제정 단계뿐만 아니라 국제적

담론 형성 단계에서도 매우 중요하다. 이에 따라 정부는 G7 또는 G20 정상회담과 같은 포럼과 경제협력개발기구(OECD) 같은 정부 간 플랫폼을 활용하여 모범 사례를 공유하고, 국제적인 차원에서 인공지능과 기계학습을 지원할 수 있는 기반으로 구축하기 위한 혁신적인 정책과 법제를 마련해야 한다.

이상과 같은 기계학습 분야의 이슈와 문제해결을 위한 법적 근거 및 정책적 지원과 더불어, 기계학습을 다양한 정책 영역에 접목하여 활용하기 위해 가장 중요하고 필수적인 것은 데이터라고 할 수 있다. 1차 산업혁명이 석탄과 증기기관, 2차 산업혁명이 석유와 전기, 3차 산업혁명이 IT시스템을 핵심적인 기반으로 하고 있다면, 4차 산업혁명은 '데이터'라는 자원을 기반으로 작동하는 생태계를 의미한다고 할 수 있다(정용찬, 2017). 이처럼 4차 산업혁명 시대의 핵심 자원은 데이터이다. 이러한 경향이 바로 OECD에서 언급한 데이터 주도 혁신(DDI: Data-Driven Innovation)이며 동시에 데이터 주도 정책(data-driven policy)의 개념이라고 할 수 있다(OECD, 2014).

이와 관련하여 보건복지 분야에서도 데이터 주도 혁신 및 정책 추진[10]을 통한 새로운 가치 창출의 필요성이 대두되고 있다(최현수, 오미애, 2017a, p. 1). 보건복지 분야에서도 다양한 종류의 데이터를 연계 및 분석하여 미래를 예측하고 새로운 가치를 창출하며, 특히 데이터 주도 정책 추진 기반을 구축하고 활용하는 것이 매우 중요하다. 과거 데이터 분석은 다양한 사회현상에 대한 단순한 정보를 제공하거나 원인을 진단하는 것이었다면, 4차 산업혁명 시대에는 미래에 대한 예측을 통하여 최적화(optimization)된 해법을 도출 및 제공하는 것이 중요하다고 할 수 있다. 보건복지 분야에서 구축되는 다양한 데이터를 기반으로 보건의료 빅데이

10) 데이터 주도 정책은 최현수, 오미애(2017a)의 일부내용을 수정·보완하여 재구성함.

터와 사회보장 빅데이터를 구축하여 강화된 개인정보보호 기술을 적용하여 적절한 범위에서 데이터를 공개하고, 기계학습 방법론을 활용한 예측을 통해 시간과 공간, 그리고 인간에 대해 최적화된 문제 해결 방안을 도출함으로써 데이터 주도 정책은 우리 삶의 다양한 분야에서 전반적으로 활용하고 새로운 경제적·사회적 가치를 창출할 수 있다.

최근 이러한 데이터 주도 정책 추진과 관련해 다양한 데이터를 연계, 구축하여 보건복지 수요에 대한 맞춤형 서비스를 제공하기 위한 연구와 실천이 공공과 민간 영역에서 다양한 형태로 진행되고 있다. 보건의료 분야에서는 우선적으로 진료 정보 등 다양한 데이터를 연계하기 위한 표준화와 분산된 데이터를 공유하기 위한 플랫폼 구축이 추진되고 있다. 또한, 건강을 중심으로 한 데이터 구축과 기술 혁신은 건강 정보에 대한 접근성과 관심을 높이고 있다. 특히, 기술적인 발전에 따라 일상생활의 다양한 측면에 대해 데이터를 광범위하게 구축할 수 있게 되었는데, 이는 데이터 주도하에 건강 문제에 대한 혁신적 정책과 솔루션을 제시하고 활용하는 것을 가능하게 만들어 국가 및 지역사회, 그리고 개인의 건강과 삶의 질을 획기적으로 향상시킬 수 있는 잠재력을 지니고 있다.

사회복지 분야에서도 빅데이터를 기반으로 기계학습을 활용한 예측과 맞춤형 서비스 중심의 지능정보사회형 혁신적 복지전달체계 구축이 시도되고 있다. 보건복지부와 한국보건사회연구원, 사회보장정보원, 서울대학교 통계학과가 협업하여 추진하고 있는 복지 사각지대 발굴 시스템은 빅데이터와 기계학습 방법을 활용하여 복지 사각지대 위험을 예측하고 이를 기반으로 발굴 대상을 추출하여 정보를 제공함으로써 현장 전문가(사회복지공무원 및 사례관리전문가)가 가구를 방문 상담하여 찾아주는 복지서비스를 제공하도록 지원하고 있다. 이는 빅데이터 구축과 기계학습 알고리즘 활용에 따른 예측 및 추천 시스템과 현장 전문가의 융합을

통한 데이터 중심 정책 추진 및 사회적 가치 창출의 대표적 사례이다. 또한 이를 기반으로 빅데이터 연계와 인공지능 활용을 통한 복지 대상의 판별 및 위험 요인 예측 결과에 따른 다양한 보건복지서비스 추천 시스템과 신청이 연계된 찾아 주는 복지 패러다임으로의 전환을 모색할 수 있다.

이처럼 보건복지 분야에서 기계학습 방법론을 활용한 데이터 주도(data-driven) 또는 데이터 중심으로 보건복지 정책을 성공적으로 추진하기 위한 세 가지 발전방향을 제안하고자 한다.

먼저, 데이터 주도 정책을 추진하기 위한 데이터 거버넌스와 인프라 구축이 선행되어야 한다. 이를 통해 미래 경쟁력의 원천인 데이터 자원의 가치 창출과 지능정보기술 활용 기반을 마련하고, 데이터와 서비스 중심의 네트워크 환경을 구축해야 한다. 최근 4차 산업혁명 시대에 대한 관심과 대응을 논의하는 과정에서도 공공과 민간 모두 4차 산업혁명 시대의 핵심적 기반이 되는 양질의 데이터가 축적되지 않거나 공유되지 않아 실질적으로 분석에 활용해 정책을 설계하거나 부가가치를 창출하는 데 한계가 있다는 문제점이 지적되었다. 보건복지 분야에서도 데이터 주도 정책을 추진하기 위해 보건의료 빅데이터와 사회보장 빅데이터 구축뿐만 아니라 보건복지 분야 데이터 상호 연계를 위한 거버넌스 체계 구축, 더 나아가 고용, 주거 등 다양한 삶의 영역으로 확장하여 데이터를 연계, 구축하는 것이 매우 중요하다고 할 수 있다. 또한, 데이터 거버넌스 또는 인프라 구축 초기부터 데이터 융합 플랫폼 구축과 데이터 매핑에 대한 계획 및 적극적 투자가 필수적이라고 할 수 있다. 이를 위해 비식별화 방법론 등 개인정보보호 기술을 강화하고 이를 기반으로 정책연구 목적으로 비식별화된 개인정보의 활용이 촉진되도록 개인정보보호 관련 법·제도 인프라를 개선해야 한다. 특히, 비식별화 조치를 하면 개인정보 수집 목적 이외에도 활용할 수 있도록 허용하되 전통적 개인정보 이외에 새로운 유

형의 개인정보보호 체계 및 기술의 구축도 필요하다.

다음으로, 데이터 중심 정책 추진을 위해서는 데이터 축적 시간과 데이터 연계를 위한 다양한 방법에 대한 고민이 필요하다. 데이터의 확충과 구축, 품질 제고와 데이터 간 연계 활용은 절대로 단시간에 이루어질 수 있는 것이 아니라는 인식을 정부 관계자 및 각 분야에서 공유하는 것이 중요하다. 초기 데이터 매핑 이후 실제로 수많은 데이터가 축적되고 이를 활용해 분석할 수 있는 데이터로 정제되기까지, 그리고 분석 결과를 기반으로 정책적 의사결정을 내리는 단계까지 축적의 시간과 기다림이 필요하다는 공감대가 필요하다. 이는 일상생활에서 빠르게 대규모로 구축되는 빅데이터와는 달리 정책영역과 연계된 데이터, 특히 사회정책 영역에서 생산 및 축적되는 데이터의 속성과 데이터 가치의 순환 주기에 대한 인식과 공감대를 바탕으로 해야 가능한 것이다. 또한, 모든 공공데이터를 원칙적으로 공개하도록 하고, 비공개 대상 공공데이터에 대해 주기적으로 공개 필요성을 검토하며, 프라이버시의 침해 없이 데이터 수집과 축적이 활성화되도록 정보 유형별(일반 정보, 비식별 정보, 개인정보) 차별화 전략을 마련하는 것도 중요하다. 이러한 과정에 데이터 기반을 구축하고 데이터 중심 정책을 추진하는 핵심 주체인 정부와 지방자치단체의 의지와 역할은 4차 산업혁명 시대의 성패를 결정하는 중요한 요소라고 할 수 있다. 데이터 주도의 정책 활성화를 위해서 정부의 역할이 매우 중요하며, 데이터 공급자 역할뿐 아니라 선도자, 촉매자, 활용자 역할 모두를 적절하게 담당해야 한다. 중앙정부와 공공기관, 지자체, 민간이 보유하고 있는 데이터의 축적과 공개를 위한 선도적 역할과 함께 데이터 생태계 활성화를 위해서 촉매자 역할을 하면서도 실질적으로 활용 가능한 데이터가 제대로 구축될 때까지 충분히 지원하고 기다릴 수 있는 문화는 정부가 주도적으로 조성해야 할 것이다. 또한, 최근 국가통계에 비해 상대적으로

열악한 지역통계를 발전시키기 위해서는 지방분권을 강화하고, 4차 산업혁명 시대에 대비해 지역통계 발전에 필요한 기술 지원과 협업, 통계 관련 지식정보의 교환·공유, 통계 개발과 활용, 데이터센터 활성화, 정책 수립을 위해 다양한 행정 자료와 통계 데이터베이스(DB)의 활용을 확대해야 한다. 데이터 기반으로 4차 산업혁명 시대에 필요한 지역 정책을 과학적으로 추진하고자 노력하는 다양한 지자체의 사례는 의미 있다고 할 수 있으며 이를 적절하게 지원하는 것은 중앙정부의 역할이다.

마지막으로, 우리 삶에 획기적인 변화를 가져올 수 있는 다양하고 새로운 가치를 창출하기 위해 데이터 주도 정책 추진과 활용에 대한 지원과 노력이 필요하다. 데이터는 기계학습의 중요한 자원이지만 이러한 자원으로서 데이터의 중요성뿐만 아니라 이를 활용한 맞춤형 정책과 서비스 개발의 중요성 또한 매우 크다고 할 수 있다. 4차 산업혁명 시대의 데이터 기반 혁신과 정책 추진에서 인공지능과 빅데이터 분석 기술의 활용 강화는 필수적이다. 현재 우리나라는 데이터의 부재라는 문제도 안고 있지만 데이터의 분석, 활용 면에서 경쟁력이 높다고 말하기 어렵다. 따라서 축적의 시간을 거쳐 방대한 데이터가 수집, 구축된 이후 데이터를 기반으로 어떻게 정책을 설계, 추진하고 혁신 전략을 마련할 것인지가 매우 중요하다. 데이터가 부가가치를 창출하는 새로운 자산으로 부각되는 4차 산업혁명 시대에 국가 경쟁력 및 우리 삶의 질과 밀접한 보건복지 분야 발전의 원천은 데이터의 공유와 활용에 있다고 해도 과언이 아니다. 영국, 미국 등 주요 국가에서 최근 데이터 경쟁력 강화 전략에 주목하는 이유는 데이터를 경제성장뿐 아니라 일자리 창출과 사회 발전을 위한 필수 자원으로 인식하기 때문이다(European Commission, 2017). 데이터 기반의 과학적 정책 결정이 이루어지는 예측형, 예방형 데이터 주도 보건복지 정책을 추진하기 위해서는 정부를 포함한 보건복지 분야 전반에서 데이

터 중심의 의사 결정 문화로 변화할 필요가 있다. 특히 보건복지 관련 빅데이터 생산·관리 인프라를 기반으로 다양한 정책 변화에 따른 국민의 삶의 질 변화와 관련된 다양한 데이터를 구축하고 이러한 데이터를 활용한 예측 결과에 기반을 둔 정책 목표 설정과 정책 설계, 환류 데이터 분석을 통한 정책 수정과 실천이 병행되어야 한다. 또한, 보건복지 분야의 데이터 활용도 제고를 위해서는 데이터 거버넌스와 데이터 공유 플랫폼에 대한 정부 지원과 사회적 신뢰 기반 구축이 필수적이다. 공공과 민간 데이터가 모두 거래되는 데이터거래소, 데이터 분석과 활용을 지원하는 데이터처리지원센터 등을 설치하여 데이터 기반의 지능형 의료서비스 제공과 지능정보사회의 사회안전망 강화를 적극 추진해야 한다.

이러한 데이터 주도 정책이 뒷받침된다면, 기계학습 분야는 더욱 발전하고 보건복지 분야에서도 이를 활용하여 다양한 정책을 추진함으로써 4차 산업혁명 시대를 맞이하고 있는 우리의 삶에 긍정적인 변화를 가져오고 중요한 영향을 미칠 수 있을 것이다.

참고문헌

강대기. (2016). 딥러닝 기반 기계학습 기술 동향. 주간기술동향, 15-16.

고학수. (2017). Policy issues surrounding artificial intelligence. algorithms & Privacy, pp. 77-78.

구글 스토어. (2017). http://store.google.com 파파고(papago) 상세 설명 이미지 (2017. 11. 29. 인출).

김은하, 추병주, 최현수, 오미애, 김용대 등. (2015). 사회보장정보시스템을 활용한 복지 사각지대 발굴방안 연구. 서울: 사회보장정보원.

김의중. (2016). 알고리즘으로 배우는 인공지능, 머신러닝, 딥러닝 입문. 서울: 위키북스. p. 77.

김인중. (2016). 기계학습의 발전 동향, 산업화 사례 및 활성화 정책 방향-딥러닝 기술을 중심으로. 소프트웨어정책연구소 이슈리포트, 2015(17), p. 34.

김정형. (2013). 컴퓨터와 인간의 게임 대결, 어디까지 왔나. Premium Chosun. http://premium.chosun.com/site/data/html_dir/2013/12/26/2013122602309.html(2017. 11. 29. 인출).

박창이, 김용대, 김진석, 송종우, 최호식. (2011). R을 이용한 데이터마이닝. 서울: 교우사.

보건복지부. (2017. 1. 9.). 보건복지부 2017 주요업무 계획 보도자료. p. 5.

성호철. (2017. 5. 10). 네이버·카카오, 인공지능에 사활 거는 까닭은. 조선비즈. http://biz.chosun.com/site/data/html_dir/2017/05/10/2017051000059.html(2017. 11. 29. 인출).

유진상. (2015. 8. 6.). 이제는 인공지능시대...ICT 기업들 경쟁 점화. 조선비즈. http://it.chosun.com/news/article.html?no=2805202(2017. 11. 29. 인출).

유재준. (2017). 초짜 대학원생 입장에서 이해하는 Generative Adversarial Nets. http://jaejunyoo.blogspot.kr/2017/01/generative-adversarial-nets-1.html?m=1(2017. 11. 29. 인출)

이태진, 정해식, 최현수, 고제이, 김현경, 신정우 등. (2016). 통계로 보는 사회보장 2016. 세종: 보건복지부, 한국보건사회연구원.

인하대 임베디드시스템공학과 블로그. (2017). http://blog.naver.com/kyb7902?Redirect=Log&logNo=221018516377 Microsoft ResNet 설명 자료 (2017. 11. 29. 인출).

정용찬. (2017). 4차 산업혁명 시대의 데이터 경제 활성화 전략. 서울: 정보통신정책연구원.

중앙일보. (2017. 10. 19.). 알파고 업그레이드…더 이상 인간의 기보 입력 안 한다. http://news.joins.com/article/22026687(2017. 11. 29. 인출).

최대우. (2017). 4차 산업혁명 시대의 빅데이터 그리고 AI. ICT Convergence Korea 2017 특강자료. 서울.

최성준, 이경재. (2017). 알파고를 탄생시킨 강화학습의 비밀. 카카오AI리포트.

최현수, 오미애. (2017a). 4차 산업혁명에 대비한 보건복지 분야 데이터 주도 정책 추진 필요성과 방향. 보건복지포럼, 8월호. 세종: 한국보건사회연구원.

최현수, 오미애. (2017b). 4차 산업혁명 및 지능정보사회의 새로운 사회적 위험과 복지 패러다임 전환 필요성. 보건복지 ISSUE & FOCUS, 제333호. 세종: 한국보건사회연구원.

한국컨텐츠진흥원. (2016). 딥러닝, 어디에 적용되고 있나. CT문화와 기술의 만남, pp. 12-15.

허핑턴포스트. (2014. 3. 21.). 페이스북, 얼굴인식 기능 '딥 페이스' 개발. http://www.huffingtonpost.kr/2014/03/21/story_n_4996658.html(2017. 11. 29. 인출).

Amazon. (2017). http://amazon.com/ 제품 설명 화면(2017. 11. 29. 인출).

Andrej, K. & Li, F. F. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. *CVPR 2015 paper*.

Apple. (2011). https://www.apple.com/kr/ios/siri/(2017. 11. 29. 인출).

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, Annals of Satistics, 24, 2350-2383.

Britton, J., Shephard, N., & Vignoles, A. (2015). *Comparing Sample Survey Measures of English Earnings of Graduates with Administrative Data during the Great Recession*. Institute for Fiscal Studies. London.

Castrounis, A. (2016). Artificial Intelligence, Deep Learning, and Neural Networks, Explained. https://www.kdnuggets.com/2016/ 10/artificial-intelligence-deep-learning-neural-networks-explained.html(2017. 11. 29. 인출).

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011a). How does your kindergarten classroom affect your earnings? evidence from project star. *Q. J. Econ*. 126, 1593-1660.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011b). The Long-term Impacts of Teachers: Teacher Value-added and Student Outcomes in Adulthood. *National Bureau of Economic Research*.

Colah. (2015). *Understanding LSTM Networks*. http://colah.github.io/ posts/2015-08-Understanding-LSTMs(2017. 11. 29. 인출).

Constine, J. (2016). *Robinhood lets china trade US stocks free through Baidu's finance app*. https://techcrunch.com/2016/06/01/ robinhood-china/(2017. 11. 29. 인출).

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297.

Domingos, P. (2012). A Few useful things to know about machine learning. New York: *Communications of the ACM*, Volume 55(10), 78-87.

EUROPEAN COMMISSON. (2017). COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS on the Mid-Term Review on the implementation of the Digital Single Market Strategy.

Fallon, B., Filippelli, J., Black, T., Trocmé, N., & Esposito, T. (2017). How Can Data Drive Policy and Practice in Child Welfare? Making the Link in Canada. *International Journal of Environmental Research and Public Health*, 14, p. 1223.

Frans, K. (2016). *Variational Autoencoders Explained*. http:// kvfrans.com/ variational-autoencoders-explained/(2017. 11. 29. 인출).

Freund, Y. & Schapire, R. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of computer and system science*, 55, 119-139.

Friedman, J. H. (2001) *Greedy Function Approximation : A Gradient Boosting Machine, Annals of Statistics,* 29, 1189-1232.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). *Additive logistic regression: a statistical view of boosting, Annals of Statistics*, 28, 337-407.

Goodfellow, I., Pouget-Abadie, J., & Mirza, M. (2014). Generative Adversarial Nets. *NIPS 2016 Tutorial*. Canada.

Gupta, S., & Karandikar, A. M. (2015). Diagnosis of Diabetic Retinopathy using Machine Learning. *J Res Development*, 3: 127. doi: 10.4172/2311-3278.1000127

Hassouna, M., Tarhini, A., & Elyas, T. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. *International Business Research*, Vol 8(6), 224-237.

He, K. (2015). Deep Residual Learning. *Conference on Computer Vision and Pattern Recognition*. Boston.

Heisler, Y. (2016. 7. 7.). http://bgr.com/2016/07/07/tesla-autopilot-software-model-x-crash-nhtsa/(2017. 11. 29. 인출).

Hinton, G. E. & Salakhutdinov, R. R. (2006). *Reducing the Dimensionality of Data with Neural Networks. Science*, 2006, Jul 28: 313.

Ian, H. W., & Eibe, F. (2016). *Evaluation-Lift and Costs*.

Intel Nervana. (2017). http://intelnervana.co/(2017. 11. 29. 인출).

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. I*CML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning*, Vol(48), 1558-1566. New York.

NVIDIA. (2017). http://kr.nvidia.com/object/drive-px-kr.html(2017. 11. 29. 인출).

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamental of machine learing for predictive data analytics-algorithms, worked examples and case studies*. London: The MIT Press.

Klein, D., & Abbeel, P. (2012). *UC Berkeley CS188 Intro to AI-Course Materials(video).*

Kotu, V. & Deshpande, B. (2014). *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. USA: Morgan Kaufmann.

Krizhevsky, A., Sutskever, I., & Hilton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *NIPS'12 Proceedings of the 25th International conference on Neural*

Information Processing Systems, Vol(1), 1097-1105.

Kuhn, M. & Johnson, K. (2013). Applied Predictive Modeling. USA: Springer Science & Business Media.

O'Brien, D., Sampson, R., & Winship, C. (2015). Ecometrics in the age of big data. Measuring and assessing "broken windows" using large-scale administrative records. *Sociol. Methodol.* 45, 101-147.

OECD. (2014). *Data-driven Innovation for Growth and Well-being.* http://www.oecd.org/sti/ieconomy/data-driven-innovation.htm (2017. 11. 29. 인출)

Raschka, S.(2017). *Lift Score.* https://rasbt.github.io/mlxtend/user_guide/evaluate/lift_score/(2017. 11. 29. 인출).

Sayama, H. (2015). Introduction to the modeling and analysis of complex systems. *OPEN textbooks SUNY.* pp. 21-22.

Science X Network. (2015). http://phys.org(2017. 11. 29. 인출).

Tensorflow blog. (2017a). 머신러닝이란. https://tensorflow.blog/%ed%95%b4%ec%bb%a4%ec%97%90%ea%b2%8c-%ec%a0%84%ed%95%b4%eb%93%a4%ec%9d%80-%eb%a8%b8%ec%8b%a0%eb%9f%ac%eb%8b%9d-1/ (2017. 11. 29. 인출).

Tensorflow blog. (2017b). *First Contact with TensorFlow.* https://tensorflow.blog/1-%ed%85%90%ec%84%9c%ed%94%8c%eb%a1%9c%ec%9a%b0-%ea%b8%b0%eb%b3%b8%eb%8b%a4%ec%a7%80%ea%b8%b0-first-contact-with-tensorflow/(2017. 11. 29. 인출).

Tikhonov, A. N., & Arsenin, V. Y. (1977). s*olutions of Ill-posed problems. Winston*, Washington, D.C.

Vijay Kotu & Bala Deshpande. (2014). Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner. Morgan Kaufmann.

WAVE. (2017). http://store.clova.ai/ko/product/wave 웨이브 상세 설명 이미지(2017. 11. 29. 인출).

WordPress. (2009). https://onionesquereality.wordpress.com/2009/03/22/why-are-support-vectors-machines-called-so/(2017. 11. 29. 인출).

Zou, H., & Hastie, T. (2005). Regularization and variable selection vis the elastic net. *Journal of the Royal Statistical Society,* Series B, 67, 301-320.

간행물회원제 안내

▶ 회원에 대한 특전

- 본 연구원이 발행하는 판매용 보고서는 물론 「보건복지포럼」, 「보건사회연구」도 무료로 받아보실 수 있으며 일반 서점에서 구입할 수 없는 비매용 간행물은 실비로 제공합니다.
- 가입기간 중 회비가 인상되는 경우라도 추가 부담이 없습니다.

▶ 회원종류

- 전체간행물회원 : 120,000원
- 보건분야 간행물회원 : 75,000원
- 사회분야 간행물회원 : 75,000원
- 정기간행물회원 : 35,000원

▶ 가입방법

- 홈페이지(www.kihasa.re.kr) - 발간자료 - 간행물구독안내

▶ 문의처

- (30147) 세종특별자치시 시청대로 370 세종국책연구단지 사회정책동 1~5F
  간행물 담당자 (Tel: 044-287-8157)

KIHASA 도서 판매처

- 한국경제서적(총판) 737-7498
- 교보문고(광화문점) 1544-1900
- 영풍문고(종로점) 399-5600
- 서울문고(종로점) 2198-2307
- Yes24 http://www.yes24.com
- 알라딘 http://www.aladdin.co.kr