정책자료
2020-01

기계학습을 이용한 노동시장 예측모형 탐색

방 형 준

한국노동연구원

목 차

표 목 차

그림목차

요 약

1. 서 론

본 연구는 기계학습의 원리를 이용하여 노동시장의 주요 변수인 경제활동참가율, 실업률, 고용률을 심층학습 모형으로 예측하는 것을 목표로 한다. 이때 예측을 위해 수치화된 통계 자료를 사용하는 것뿐만 아니라 여론 및 심리를 대변하는 변수로 신문 기사라는 비정형 데이터를 활용하여, 여론이 경제 변수에 영향을 미치는지를 정형 데이터만을 활용한 예측 결과와 비교하여 분석함으로써 검증하고자 한다.

2. 기계학습과 예측 모형

제2장에서는 기계학습 및 심층학습의 원리와 방법론, 실제 모형 설계 시 고려한 여러 요인을 살펴보았다. 심층학습을 이용하면 자연어 처리가 가능한데, 기존에는 컴퓨터가 바로 인식하거나 처리할 수 없었기 때문에 사용할 수 없었던 자연어의 영역은 이제 충분한 컴퓨팅 자원만 있으면 분석할 수 있게 되었다. 따라서 그간 수치화된 통계를 주로 이용하여 분석한 연구에 자연어 처리는 새로운 활력을 불어넣을 수 있을 것으로 기대된다.

본 연구에서는 예측 모형을 설계할 때 정형 데이터와 자연어로 구성된 신문 기사 제목인 비정형 데이터를 모두 사용하는 것을 고려하였다. 이러한 정형 및 비정형 데이터를 활용하여 노동시장의 주요 변수들을 심층학습 모형으로 예측하고자 한다. 예측 모형은 세 가지로, 첫 번째는 노동시장 관련 변수들만을 이용하여 노동시장 변수를 예

측하게 함으로써 노동시장에 강한 내생성이 존재하는지를 알아볼 것이다. 두 번째 모형에서는 노동시장 통계 이외에도 이용 가능한 모든 수치 데이터인 정형 데이터를 활용하여, 실물 경제의 각 분야에 대한 통계치들뿐만 아니라 각종 사회경제적 지표들, 경제주체들의 기대를 반영한 전망 지수들과 각종 인구 통계까지 사용하여 노동시장 변수를 예측함으로써 노동시장 변수들이 실물 경제에 얼마나 영향을 받는지를 첫 번째 모형과 비교하여 파악하고자 한다.

세 번째 모형은 비정형 자료인 신문 기사 제목까지 포함하여 노동시장 변수를 예측하게 하는 것으로, 여기서 사용한 비정형 자료는 지난 20여 년간의 중앙 일간지 및 주간지 신문 기사 제목 전체이다. 자연어는 기계학습을 위해 분석하기 위해서 전처리 과정을 요하는데, 연구자의 임의성과 주관을 최대한 배제하기 위해 최소한의 전처리만 수행했다.

전처리를 거친 비정형 자료는 분석의 최소 단위인 형태소로 수치 변환된 후, 최종적으로 월별 5차원 배열의 텐서로 변환된다. 자연어는 일반적으로 단어의 나열 순서나 기사의 순서가 문맥이나 경기 변동의 흐름을 설명하는 의미를 가지기 때문에 순차적인 배열에도 의미를 부여하는 순환 신경망 방법을 사용하였다. 이렇게 개별 신문 기사 제목은 일별 단위로 변환된 후, 매월마다 일별 신문 기사 제목에 대한 텐서를 병합하여 다시 월 단위 수치 벡터로 변환된다. 이렇게 일 단위 자료들을 월 단위로 변환하고 나면, 예측하고자 하는 노동시장 지표와 동일하게 모든 자료가 월 단위로 주어지고, 이를 가지고 세 개의 모형은 각각 경제활동참가율, 실업률, 고용률을 예측한다. 이때 예측값과 실제값은 차이가 있는데, 모형에 대한 평가는 두 값의 차이를 측정한 평균 제곱 오차를 계산한 다음 이를 최소화하는 방향으로 모형을 개선하도록 설계하였다.

기계학습을 통해 설계된 모형은 일반성과 적합도로 평가하는데, 일반성은 모형이 개개의 변수에 크게 영향을 받지 않으며 추세적으로 실제값을 근사하게 예측하는지에 대한 지표이며, 적합도는 주어

진 모든 설명변수를 활용하여 학습 데이터로 예측값을 최대한 정확하게 설명하는지에 대한 지표이다. 대부분의 모형에서 적합도와 일반성 간에는 상충 관계가 존재하므로 이 두 지표를 모두 적절히 만족하는 최적의 모형을 찾기 위해 모든 기계학습 모형에 대해서 drop-out rate과 조기 종료를 설정한다. drop-out rate이란 모형의 일반성을 높이기 위해서 모형을 설계한 후 임의로 일정한 비율의 설명변수를 삭제하고 모형의 성과를 평가하도록 하는 지표로, drop-out rate이 높으면 높을수록 모형의 일반성은 높아지지만 적합도가 낮아지고, drop-out rate이 낮으면 적합성이 높아지는 반면 과거의 데이터만을 잘 설명하게 되는 과적합의 문제가 발생할 수 있다. 한편 학습을 통해서 개선되는 모형들이 일정 개수 기존 모형보다 성과가 떨어지면 미리 설정한 stopping point가 충족되면 기계학습은 최적 모형을 결정하고, 해당 모형에 기초하여 노동시장 변수들을 예측한다.

3. 기계학습 모형의 예측 결과

본 장에서는 세 가지 기계학습 모형을 설계한 후, 각 모형이 경제활동참가율, 실업률, 고용률을 예측케 하였다. 첫 번째 모형은 노동시장과 직접적으로 관련된 변수의 시계열 자료만을 이용하여 세 가지 변수를 예측하게 하였다. 이 모형에서는 따라서 노동시장에 존재하는 계절성과 내생성, 그리고 노동시장의 자체적인 장단기 추세가 주요한 분석 대상 변수로 들어갈 것이다. 두 번째 모형은 실물 경제의 상황과 경기 현황 및 전망, 수출입 등을 대변하는 다양한 정형 데이터를 이용하여 노동시장 변수를 예측하였다. 세 번째 모형에서는 두 번째 모형에서 사용한 정형 데이터 전체에 지난 20년간의 신문기사 제목이라는 비정형 데이터까지 포함하여 노동시장 변수들을 예측케 하여, 노동시장의 주요 변수들이 여론이나 심리의 영향을 받는지 여부를 파악하고자 했다. 분석 결과는 다음과 같다.

경제활동참가율의 경우, 평균 제곱 오차로 측정한 모형의 예측력

에 있어서는 모형 I이 가장 나은 결과를 보여주었으며, 모형 III이 가장 부정확한 예측 결과를 나타냈다. 반면 경제활동참가율의 증감 추세를 얼마나 정확하게 예측했는지도 고려했을 때 모형의 예측력은 단정 지어 말하기 어려운 결과가 나왔다. 실업률과 고용률에 있어서는 모형 I이 평균 제곱 오차나 증감 여부 모두에서 가장 나은 것으로 나타났다.

이러한 결과를 통해 다음과 같은 사실을 알 수 있다. 첫째로 노동시장은 내생성이 강하여 실물 경제의 움직임에 즉각적으로 반응하지 않는 경직성을 어느 정도 가지는 것으로 보인다. 여타 실물 변수를 포함한 모형 II의 예측력이 모형 I에 비해서 전체적으로 좋지 않게 나왔기 때문이다.

노동시장의 주요 변수들을 예측하는 데 있어 신문 기사 제목은 큰 영향을 미치지 못하는 것으로 나타나, 심리나 여론 등이 노동시장에 큰 영향을 주지 않는 것으로 보인다. 이는 신문 기사 제목이라는 비정형 데이터까지 포함한 모형 III이 실업률과 고용률에서는 예측력이 가장 좋지 않았으며, 경제활동참가율에 대해서도 가장 좋은 예측력을 보이지는 않았기 때문이다.

이러한 결과가 나타난 이유로는 노동시장 이외의 각종 실물 경제 지표와 신문 기사가 노이즈처럼 작용하여 예측력을 떨어뜨리기 때문인 것으로 보인다. 기계학습에서는 학습 데이터의 크기가 클수록 모형의 설명력과 예측력이 올라가지만, 여기서는 노동시장과 무관한 변수들까지 설명변수에 삽입되어 모형이 과적합한 것으로 보인다.

그러나 여기서의 분석 결과를 통해 노동시장이 사람들의 심리나 여론으로부터 전혀 영향을 받지 않는다고 해석할 수는 없다. 이러한 결과가 나타난 이유는 노동시장과 무관한 여러 신문 기사나 사건들이 포함되었기 때문일 수도 있으며, 자연어를 처리하는 방법을 바꿀 경우에도 모형의 예측력이 바뀔 수 있기 때문이다.

4. 결 론

본 연구에서는 노동시장의 주요 변수들인 경제활동참가율과 고용률, 실업률을 기계학습과 심층학습을 통해 예측해 보았다. 이를 통해 노동시장은 내생성이 강하여 노동시장 변수만을 사용한 모형이 가장 높은 예측력을 가진 것으로 나타났으며, 여론 및 사람들의 심리를 대변하는 신문 기사 제목까지 사용한 모형의 예측력이 가장 낮은 것으로 나타났다.

그러나 이러한 예측 결과를 바탕으로 노동시장의 주요 변수들은 노동시장의 내적 요인들로만 설명하는 것으로 충분하다고 생각하는 것은 성급한 결론이며, 분석 결과에서는 학습 데이터의 부족으로 인해 모형이 과적합했을 가능성을 배제할 수 없다. 그러므로 시간이 흐르면서 학습 가능한 데이터가 누적될 경우 연구 결과는 언제든지 바뀔 수 있으며, 기계학습의 방법이나 변수 선택에 따라서도 결과는 달라질 수 있다. 그러나 과거 20여 년간의 데이터를 이용하여 다음과 같은 결론을 얻었다.

첫째, 경제 주체들이 노동 공급이나 노동 수요에 대한 의사결정을 할 때에는 노동시장의 상황을 가장 많이 고려하고, 실물 경제도 일부 고려하지만 여론이나 언론의 경향에는 많은 영향을 받지 않을 수 있다. 이러한 결과가 나오는 한 가지 이유로, 신문의 경우 논쟁적이거나 사람들의 주목을 받는 주제들이 과대 대표되는 반면 노동시장에 많은 영향을 미친다 하더라도 주목받지 못하는 주제는 과소 대표될 수 있기 때문이다.

둘째, 경제 변수를 예측할 때에는 노이즈가 포함되지 않도록 변수 선택에서부터 많은 노력을 기울여야 한다는 점이다. 일반적으로 기계학습은 데이터의 양이 많아질수록 성능이 나아지지만, 다수의 데이터가 종속변수와 무관하다면 노이즈에 의해서 오히려 성능이 떨어질 수 있다.

셋째, 노동시장 변수를 포함한 여러 경제 변수들을 예측하는 데 있

어서는 계절성과 내생성이 가장 적절한 설명변수일 가능성이 높다는 점이다.

넷째, 세 가지 모형 모두 실제값을 예측하는 데 있어 좋은 성과를 보여주지 못한 점을 통해 학습 데이터의 양이 중요하다는 것을 확인했다. 학습 데이터의 숫자가 적은 경우 모형은 필연적으로 과적합하여 예측력이 떨어질 수 있다.

마지막으로, 기계학습은 모형의 결과에 대해서 어떠한 설명도 제공하지 않고 단순히 결과만을 제공하므로, 해석에 신중을 기해야 하며, 연구의 설계 과정에서부터 해석 가능성을 염두에 두고 기계학습을 수행해야 한다.

제 1 장
서 론

제1절 문제의식

최근 빅데이터와 기계학습(machine learning) 및 심층학습(deep learning) 등에 대한 세간의 관심이 고조되고 있다. 이러한 관심은 비단 컴퓨터공학을 비롯한 유관 학계에만 한정되지 않고 정보통신업종 등을 비롯한 관련 산업을 넘어 전통적인 농업과 제조업 및 공공행정 분야까지 망라하고 있으며, 아울러 빅데이터나 기계학습 전문가들뿐만 아니라 휴대전화에서 앱을 이용하고 전자상거래를 활용하는 개인에 이르기까지 분야와 직업을 불문하고 널리 퍼져 있다. 각계각층에서 관심을 가진다는 것은 한편으로는 빅데이터에 대한 활발한 논의의 장이 열리고 빅데이터 산업 및 기계학습 관련 연구를 지원하기 위한 다양한 방법을 사회 전반에서 모색하는 계기가 되기도 하지만, 빅데이터나 기계학습에 대한 통일된 정의가 부재한 상황에서 각자 나름의 정의를 가지고 백가쟁명 하다보니 합의된 의견을 도출하기 힘들며, 산발적으로 논의가 전개되는 일도 벌어지고 있다.

그러나 빅데이터가 정확하게 무엇인지에 대해서는 심지어 학계에서도 명확한 정의를 내리지 못하고 있다. 오히려 빅데이터의 성격을 고려한다면 명확한 정의를 찾고 규정하는 것이 불가능에 가까운 일일 것이다. 이렇게 명확한 정의가 부재(不在)하는 이유는 기술 진보에 따라 컴퓨팅 시스템이 분석할 수 있는 데이터의 범위가 점차 넓어지고, 일정한 시간 내

에 처리할 수 있는 데이터의 양도 증가하고 있기 때문이다.

빅데이터에 대한 정의는 방형준・손연정・노세리(2019)에서도 다루고 있으며 빅데이터의 특성상 언제 어디서나 통용되는 명확한 정의를 찾는 것은 불가능하므로, 본 장에서는 자연어(natural language)가 빅데이터에 속하는지 여부에 대해서 심도 있게 다루지 않을 것이다. 다만 방대한 자연어 자료는 일반적으로 전통적인 3s(conventional 3Vs)나 추가적인 Vs(additional Vs)에 부합하는 경우가 많다는 사실만 밝혀두고 넘어간다.

빅데이터에 대한 용어가 본격적으로 쓰이기 시작한 것은 1990년대로, 당시에는 컴퓨팅 기법 및 데이터 처리 방식으로 분석하는 데 오랜 시간이 소요되는 데이터를 빅데이터라 지칭했다. 하지만 컴퓨팅 관련 하드웨어 및 소프트웨어의 발전으로 20~30여 년 전 빅데이터라 지칭되었던 수준의 데이터들은 더 이상 빅데이터의 정의에 부합하지 않게 되었고, 그때에는 수치적인 분석이 쉽지 않아 분석 가능한 데이터에서 제외하곤 했던 문자, 사진, 영상 등도 이제는 처리 가능한 데이터의 영역으로 진입했다.

최근 빅데이터 관련 학계에서 주목하는 주제 중 하나는 자연어 처리(natural language processing)이다. 자연어란 우리가 일상에서 사용하는 문자 및 언어를 지칭하는 것으로, 기계어와 대립되는 개념이라 생각하면 된다. 기계어는 컴퓨터가 바로 읽어낼 수 있는 형태의 언어를 지칭한다. 하지만 우리가 일상에서 사용하는 언어는 기계어와 상이하기 때문에 과거에는 자연어를 컴퓨터가 처리하는 것은 상상하기 힘들었다. 단순히 기계어를 입·출력하는 것은 이미 수십 년 전부터 있었으나, 자연어를 컴퓨터가 해석하고 분석하는 것은 최근에서야 가능해진 일이다. 자연어 처리가 가능해지면서 컴퓨터로 분석 가능한 데이터의 양도 폭발적으로 증가하였다. 과거부터 축적되었던 인류의 대부분의 지식과 정보는 문서나 그림, 음성 등의 형태로 저장되었기 때문이다.

기계어와 자연어의 경계를 허물려는 시도는 이미 오래전부터 있었다. 그러나 20여 년 전까지만 해도 기계어를 이해하고 입력하기 쉽게 자연어로 바꾸려는 노력이 더 많았다. 즉, 기계어를 직접 사용하는 것은 학습하기도 어렵고 실사용에도 애로사항이 많았기 때문에 최대한 자연어에 가깝게 프로그래밍 언어를 만들어서 사람들이 보다 쉽게 컴퓨터를 사용하

고 이해할 수 있도록 만들려 했다. 이를 위해 등장한 것이 컴파일러로, 컴파일러는 사람이 입력한 프로그래밍 언어를 기계어로 번역하는 장치이다.

이처럼 기계어를 자연어 형태로 바꾸려는 노력은 예전부터 있어 왔으나, 자연어를 기계가 처리하고 분석 가능케 하는 작업은 최근에 와서야 가능해졌다. 그전에도 문서 작업 등을 컴퓨터를 통해 처리하였으나, 이는 엄밀히 말하여 컴퓨터가 자연어를 처리했다기보다는 컴퓨터가 사람의 명령에 따라 자연어를 표시하는 단순한 입출력에 불과했다. 컴퓨터가 자연어를 직접 읽고 이해하며 처리하는 것은 1990년대에서야 비로소 초보적인 수준에서 가능해졌다.

이러한 기술 진보는 비단 컴퓨터의 하드웨어 성능이 개선된 것뿐만 아니라 소프트웨어적인 발전도 큰 몫을 하였다. 특히 기계학습과 심층학습은 자연어를 처리하게 하는 데 있어 촉매와 같은 역할을 하였다. 컴퓨터가 자연어를 분석할 때 어려운 점은 자연어의 모든 경우에 대해서 사람이 일일이 학습을 시키기 어려웠기 때문이다. 또한 사람마다 사용하는 자연어의 패턴이나 어휘가 다르므로 이것이 같은 의미를 지칭하는지 아닌지를 판단하는 것 역시 심층학습 없이는 불가능하였다. 심층학습은 컴퓨터에게 일정한 양의 학습할 데이터를 제공하기만 하면 시스템이 주어진 데이터 내에서 일정한 패턴을 찾은 후, 데이터를 통해 탐색한 패턴이 데이터에 부합하는지를 스스로 점검하면서 학습한다.

기계학습은 비단 이러한 자연어 처리뿐만 아니라 기존에 일상에서 널리 사용하던 수치화된 자료를 분석하는 데에도 새로운 장을 열었다. 과거에는 경제현상을 분석하거나 예측하기 위한 모형을 설계할 때 주로 선형회귀모형이나 다중회귀모형을 사용하였다. 하지만 기계학습을 통해 경제학자들은 라쏘(Lasso)나 릿지(Ridge) 등을 활용하여 보다 많은 제약조건을 충족시키는 변수나 모형을 탐색하는 것이 가능해졌다. 이때 많은 기계학습 모형에서 과적합(over-fitting)이 발생할 수 있으나, 라쏘 등은 모형에서 일정한 페널티 효과를 넣어 과적합을 방지하고, 이를 통해 주어진 변수들을 이용하여 보다 해석력이 좋은 일반화된 모형을 탐색하는 것을 가능케 한다. 여전히 기계학습이나 이를 이용한 라쏘에서 해석상에 어려움이 없는 것은 아니나, 적어도 적합도나 설명력이 높은 모형을 찾는다는

목표에는 한 걸음 더 다가가게 되었다.

본 연구는 여기서 한 발 더 나아가, 일반적인 경제학의 분석 방법에서 보다는 예측력을 더 높이기 위해 탐색 모형에 매우 높은 수준의 자유도를 부여한 후, 이에 기초하여 지도 학습(supervised learning)으로 노동시장의 주요 변수들을 예측하고, 이를 통해 여러 경제지표를 포함한 정형 데이터(structured data)나 비정형 데이터(unstructured data)가 얼마나 효과적인지를 알아볼 것이다. 아울러 비정형 데이터를 이용한 분석 결과를 이용하여 경제 현상의 흐름을 설명하거나 예측하는 데 있어서 통계로 관찰 불가능한 요소들이 얼마나 작용하는지도 살펴볼 것이다.

이렇게 기계학습을 통해 노동시장 변수를 예측하는 것은 회귀분석을 포함한 기존의 경제학 분석과 비교하여 다음과 같은 점에서 차이가 있다. 첫째, 기존 분석에서는 모형의 선택이나 분석에 사용하는 변수를 선택하는 데 있어 연구자의 주관이나 임의성이 개입될 여지가 크다. 하지만 본 연구에서 수행할 기계학습에서는 프로그램이 예측을 위해 가장 적합한 모형과 변수들을 추출하도록 하여 이러한 임의성이나 주관을 배제할 수 있다.

둘째로 비정형 데이터를 분석에 포함시킬 수 있다는 점이다. 기존에 경제학에서 수행해 왔던 분석은 모두 정형 데이터만을 분석할 수 있었던 반면, 기계학습에서는 비정형 데이터까지 포함하여 분석함으로써 정형 데이터와 비정형 데이터 간의 관계, 비정형 데이터가 예측하고자 하는 변수와 얼마나 깊은 관계를 가지고 있는지 등도 아울러 파악할 수 있다.

마지막으로 앞서 언급한 바와 같이 모형에 거의 완전한 자유도를 부여하여 예측력을 극대화함으로써, 과연 노동시장의 변수들을 얼마나 잘 예측할 수 있는지를 측정해 볼 수 있을 것이다. 만일 기계학습 모형에서 노동시장 변수를 거의 정확하게 예측해 낸다면 해당 예측력을 가능한 최대의 예측력으로 놓고, 이러한 예측력에 근접한 결과를 내는 모형을 찾음으로써 역으로 노동시장 변수들이 가지는 특성을 유추해 볼 수 있다. 이는 아울러 기계학습 모형이 가지는 단점인 예측 결과에 대한 설명을 제공하지 않으며, 어떠한 모형으로 예측했는지를 알려주지 않는다는 단점을 극복할 수 있을 것으로 보인다. 모형에 계절성, 각종 고정효과, 자기상관 등

여러 요소를 고려하면서 기계학습 모형의 결과와 비교하여 유사한 결과를 내는 모형을 찾은 후, 모형을 통해서 기계학습의 예측 결과를 간접적으로 추론하거나, 혹은 과거 노동시장의 움직임을 설명하고 변동의 원인을 찾을 수 있다.

제2절 연구의 구성

본 연구의 목적은 가용한 국내외의 수치화된 통계 자료를 기반으로 노동시장의 주요 변수들을 예측해봄과 아울러, 자연어 처리 기법을 통해 한국어로 된 신문 기사를 활용하여 모형의 예측력이 높아지는지를 검증하는 데 있다. 이를 바탕으로, 노동시장의 주요 변수를 예측하거나 영향을 미치는 변수들은 어떠한 변수들인지 탐색함과 아울러, 신문 기사 등으로 대변되는 여론이 경제 변수에도 영향을 미치는지 알아보고자 한다. 기존 경제학 연구에서는 여론이나 심리가 경제에 미치는 영향을 파악하기 위해서 여론이나 심리를 대변한다고 보는 간접적인 통계 자료를 분석에 추가하여 왔으나, 본 연구에서는 여론의 영향을 직접적으로 모형에 추가함으로써 여론이 얼마나 노동시장에 영향을 미치는지를 보다 정확하게 측정할 수 있다.

우선 제2장에서는 기계학습의 원리와 자연어 처리를 위한 방법을 설명한 후, 본 연구에서 활용할 정형 데이터 및 비정형 데이터에 대해서 소개할 것이다. 또한 이러한 데이터를 활용해서 모형을 어떻게 설계했고 프로그램이 예측하는 과정은 어떻게 되는지 설명할 것이다.

제3장은 노동시장 주요 변수인 경제활동참가율, 고용률, 실업률을 각각 예측하는 모형을 구현한 결과를 살펴볼 것이다.

우선, 첫 번째 모형에서는 수치화된 통계 자료만을 가지고 노동시장 변수들을 예측할 것인데, 통계 자료를 다시 두 유형으로 분류하였다. 첫 번째 유형은 노동시장 변수들로, 학력수준별, 연령별, 지역별 노동시장 변수들이다. 즉, 예측하고자 하는 변수들과 가장 밀접하게 연관된 노동 통계

만을 학습해서 상술한 세 수치를 예측하는 것이 첫 번째 모형이다. 두 번째는 노동시장 변수들뿐만 아니라, 인구, 주택, 경기, 수출입, 금융시장, 산업 생산 및 물가 등등, 가용한 국내의 모든 통계 자료를 입력하여 경제활동참가율, 고용률, 실업률을 예측할 것이다. 이 두 모형의 결과를 비교함으로써 노동시장 변수들이 노동시장 내외의 실물 경제 요인들과 어떠한 관계를 가지는지 간접적으로 살펴볼 수 있을 것이다.

세 번째 모형에서는 앞선 모형에서 활용한 모든 통계 자료에 국내 신문 기사의 제목을 자연어 처리한 후 추가하여 예측 모형을 구현할 것이다. 이를 통해 수치화된 통계만을 활용한 모형의 예측력과 자연어를 포함한 모형의 예측력을 비교하여 경제 현상을 설명하거나 예측하는 데 있어 통계상으로 관찰 불가능한 요소들이 실물 경제 변수에 얼마나 영향을 미치는지에 대해서 살펴보고자 한다. 이후에는 상술한 모형의 예측 결과들을 종합하여 노동시장 주요 변수들과 밀접하게 연관되어 있는 경제 변수는 무엇인지 탐색해 보고, 경제 현상을 설명하는데 있어 고려해야 할 변수들이 무엇인지 고찰해 볼 것이다.

제4장에서는 이후 각종 경제모형을 설계하거나 예측 모형을 세울 때 기계학습이 기여할 수 있는 바가 있는지를 검토하고, 어떠한 변수들을 고려하는 것이 합리적일 수 있는지 대안을 모색하고자 한다.

제 2 장
기계학습과 예측 모형

제1절 기계학습의 원리

1. 기계학습의 정의와 원리

기계학습과 심층학습이 명확하게 구분되는 것은 아니다. 연구자에 따라서는 기계학습과 심층학습을 동의어로 사용하기도 하고, 어떤 연구자들은 심층학습을 기계학습과는 독립된 개념으로 보기도 한다. 하지만 일반적으로 심층학습은 기계학습 알고리즘의 한 종류로 이해하고 있기 때문에 본 연구에서는 기계학습의 하위 개념으로 간주하겠다. 심층학습에서 사용하는 대부분의 방법론은 기계학습의 방법론을 차용하거나 변용한 경우가 많으며, 특수한 상황에서 적용하는 기계학습에 가깝기 때문이다.

기계학습은 사전적으로 “스스로 학습하여 개선해 나가는 컴퓨터 알고리즘(the study of computer algorithms that improve automatically through experience)”[1]이다. 다시 말해서 기계학습이란 기계, 즉 컴퓨터가 스스로 주어진 데이터를 이용하여 학습한 후 데이터 내에서 일정한 규칙이나 패턴을 찾아내어 이를 다른 데이터들에 적용하는 방법이라 할 수 있다.

---

1) Mitchell, Tom(1997: preface XV), *Machine Learning*, New York : McGraw Hill.

기계학습은 크게 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 나눌 수 있으며, 최근에는 이 둘을 혼합한 반지도 학습(semi-supervised learning)도 사용하고 있다. 지도 학습은 사람이 개별 데이터 입력값마다 레이블을 부여해 주면 컴퓨터가 입력값과 레이블을 이용해서 시스템을 학습하는 것이다. 최종적으로 컴퓨터가 처리해야 하는 데이터의 양이 100만 개라 가정하면, 일일이 사람이 수작업으로 처리하는 것보다야 컴퓨터를 이용해서 단순 반복작업을 사람이 직접 수행하는 것이 시간이나 비용 측면에서 보다 절약되겠지만, 기계학습은 이러한 작업에 들어가는 노력을 보다 낮춰준다. 컴퓨터에게 학습을 위해 데이터 1만 개에 대해서 사람이 작업한 결과를 준 후, 이를 학습해서 나머지 99만 개를 처리하도록 한다면 사람이 해야 하는 업무의 양이 크게 줄어들게 된다. 또한 사람이 일일이 처리하는 경우 오류의 발생 가능성도 높으나, 기계학습으로 대체하는 경우에는 초기 입력값에 큰 오류가 없으며, 학습에 사용된 데이터들의 성격이 나머지 데이터와 전혀 다르지 않다면 작업에서 오류의 가능성도 현저히 줄어들게 된다.

그러나 여기서 지도 학습의 단점도 동시에 확인할 수 있다. 지도 학습을 위해서는 초기에 학습을 위한 데이터를 제공해야 하는데, 이 데이터의 레이블이 정확할수록 시스템이 보다 정확하고 효과적이며 신속하게 학습을 수행한다. 하지만 초기에 주어진 학습 데이터의 정확성이 떨어지거나 질이 나쁜 경우 학습의 정확성이나 효율성도 마찬가지로 저하된다. 따라서 지도 학습의 경우 초기 학습 데이터를 만드는 데 많은 시간과 노력이 소요되며, 경우에 따라서는 전문가의 개입을 필요로 하기 때문에 많은 비용이 소요될 수 있다. 또한 요구되는 레이블의 수준이나 개입 수준이 높은 경우에는 학습을 위해 많은 시간과 노력이 소요됨에 따라 최종적으로 생성되는 학습용 데이터의 양이 적어질 수도 있다.

비지도 학습은 지도 학습과는 반대로 시스템이나 컴퓨터가 스스로 주어진 데이터를 이용하여 학습하는 것이다. 이 경우, 시스템이 수행하는 학습이 어떠한 것이며 그 학습이 정확한지에 대해서 사람이 확인할 수 있는 방법이 없다. 따라서 일정한 수준 이상의 정확성을 요구하거나 주어진 문제에 대한 답을 찾아야 하는 경우에는 비지도 학습 방법을 적용할 수

없다. 비지도 학습이 활용되는 경우는 많은 양의 데이터를 준 후 이들을 군집화(clustering)하거나 혹은 집단 내에서 데이터들의 분포를 추정하는 분야 등이다.

최근 비지도 학습은 자연어 처리와 연관되어 각광받고 있다. 많은 양의 문헌이나 텍스트에서 주제어를 탐색하거나 주제어 간의 일정한 규칙이나 관계를 찾는 작업을 수행할 때 최근 비지도 학습이 많이 쓰이고 있는데, 대표적으로 잠재적 디리클레 할당(Latent Dirichlet Allocation : LDA)이나 문헌 기반 탐색(Literacy Based Discovery : LBD)을 들 수 있다.

문헌 기반 탐색은 텍스트 마이닝(text mining)에서 최근 사용되는 방법 중 하나로, 특정 주제와 관련된 문헌, 연구, 결과물, 각종 보고서 등의 문건을 수집한 후, 연구자의 관심사항과 관련된 키워드에 주석(annotation)을 달아서 관심 주제어들 간의 잠재적 (인과) 관계를 비지도 학습 형태로 제공하는 방법이다.

COVID-19와 같이 아직 백신이나 치료법이 개발되지 않은 새로운 질병에 대한 연구를 수행할 때, 연구자는 Corona Virus와 관련하여 현재까지 진행된 수많은 관련 문헌 및 자료들을 탐색해야 한다. 사람이 이 과정을 직접 일일이 수행하기에는 시간이 많이 소요될 뿐만 아니라 인지적 능력의 한계로 인해 탐색된 문헌 내에 포함된 주제들 간의 잠재적 관계(또는 새로운 연구 주제)를 파악하기 어렵다는 단점이 있다. 이 과정에서 문헌 기반 탐색을 사용하면 문헌 탐색에 소요되는 시간을 단축하고 잠재적 연구 주제에 대한 아이디어를 얻을 수 있다. 문헌 기반 탐색의 과정은 다음과 같다. 우선 COVID-19 관련 의학 논문과 보고서, 병원 차트 기록 등의 자료를 수집한다. 이후, 수집된 텍스트 중 일부에 연구자가 관심 있는 주제와 연관된 키워드들을 찾아서 주석을 부여하는데, 예를 들어 환자의 특성을 나타내는 변수 중 하나인 연령에 대해서 'age'라는 주석을 달고, 키나 몸무게, 거주지 등에 대해서도 마찬가지로 각각의 주석을 부여하며, 증상이 나타난 신체 부위나 구체적인 증상에 대해서도 각각 유형별로 주석을 제공한다. 아울러 처방한 약, 치료 방법, 효과를 보는 데 걸린 시간, 완치 여부 등등에 대해서도 통일된 주석이나 레이블을 제공하면, 프로그램은 텍스트를 인식한다. 이 과정은 얼핏 사람이 레이블이나 주석을 달아

서 제공하기 때문에 지도 학습으로 보일 수 있다. 하지만 프로그램이 주어진 초기 텍스트에서 환자의 개인 특성, 치료법이나 처방한 약, 완치 여부 등등에 대해서 주석을 통해 인식하면, 이를 바탕으로 수집된 전체 텍스트에 해당 주석을 확대해서 적용한다. 문헌 기반 탐색은 주석이 부여된 전체 텍스트에 대해 비지도 학습을 수행한 후, 결과물로 잠재적인 연구 주제 목록이나 시각화된 형태로 주제어 및 관련어에 대한 탐색 도구를 산출한다. 이를 바탕으로 COVID-19 치료에 대해서 개인 특성별로 가장 효과적이었거나 효과가 없는 치료법이나 약은 무엇인지, 어떠한 증상에 대해서 어떠한 약품이나 구성물이 효과가 있는지를 파악하여, 가장 효과적인 치료약을 선택하거나 혹은 치료약을 개발하는 과정에서 초기 오차를 줄이고 탐색 시간을 단축시킬 수 있다.

문헌 기반 탐색에서 초기에 제공하는 주석은 프로그램이 학습해야 하는 데이터라기보다는 프로그램이 이후에 유형을 분류할 때 사용하는 지침을 제공하는 것이며, 실제 분석은 관심 주제어들 간의 관계를 추출하는 것으로, 이 과정에서는 어떠한 외부 학습도 수행하지 않는다.

잠재적 디리클레 할당은 주제 분석(topic modelling) 기법 중 가장 널리 활용되는 방식으로, 텍스트 마이닝으로 제공된 문서들 내에서 특정 주제와 연관된 연관어들의 분포를 확률적 모형으로 보여주는 방법이다. 특정한 주제와 관련된 문서를 수집한 후, 해당 문서들에 등장하는 단어들의 빈도를 분석하면 특정 주제와 연관된 단어 혹은 표현이 무엇인지를 파악할 수 있다. 이러한 과정 자체는 TF-IDF(term frequency - inverse document frequency)와 크게 다르지 않다. 잠재적 디리클레 할당과 TF-IDF와의 차이는, TF-IDF의 경우에는 단어가 한 문서에 얼마나 많이 등장하며 해당 단어가 수집된 문서 혹은 주제와 연관된 문서들 중에 몇 개에 등장하는지의 단순 분포에 초점을 맞춘 것이라면, 잠재적 디리클레 할당은 문헌 내에서 해당 단어가 잠재적으로 갖는 의미, 즉 문맥상의 의미를 수치적으로 파악하기 위한 확률 모형을 찾으려고 한다는 점에 있다. 따라서 잠재적 디리클레 할당은 수집된 텍스트를 통해서 기존에 알지 못했던 주제 및 그 주제와 연관된 단어들 혹은 표현들을 보여줌으로써 하나의 현상과 연관되어 있었지만 기존에는 인식하지 못한 다른 원인이나 결과를 탐색하거

나 병행하는 현상을 찾는 것이 가능하다. 잠재적 디리클레 할당은 특히 수집된 문서들을 일정한 체계에 따라 분류한 후 해당 분류에 의해서 특정 주제가 어떠한 연관어와 관계를 가지거나 어떠한 특성을 보이는지를 분석하는 데 유용하다.

마지막으로 살펴볼 기계학습의 유형은 반지도 학습(semi-supervised learning)이다. 반지도 학습은 지도 학습과 비지도 학습의 혼합이며, 실제로는 지도 학습에 가까운데, 앞서 지도 학습을 위해서는 사람이 직접 학습용 데이터에 레이블을 부여하는 작업을 해야 함을 언급하였다. 하지만 다수의 학습 데이터에는 레이블이 부여되었으나 일부 학습 데이터에는 레이블이 누락된 경우 시스템이 레이블이 있는 다수의 데이터를 학습하여 레이블이 없는 소수의 학습 데이터를 보충한 후, 완결된 학습 데이터로 재학습을 수행한다. 그 이후 나머지 데이터들은 완결된 학습 데이터로 학습한 결과를 가지고 분석을 수행한다. 반지도 학습의 경우, 대부분 지도 학습과 크게 차이가 나지 않으며, 지도 학습으로 간주되기도 한다.

기계학습 중에서도 특히 심층학습에 대한 관심이 높아지고 있는데, 심층학습은 인공신경망(artificial neural networks)에 기반하는 기계학습 기법이라 할 수 있다. 인공신경망은 복잡한 연산을 효율적으로 수행하고 계산된 결과물의 정확성을 제고하기 위해 컴퓨터 공학에서 시도된 여러 방법론 중 하나로, 생체신경망(biological neural networks)의 구조를 모방하여 유사하게 연산을 처리하는 방식이다. 인공신경망은 이미 1960년대에 고안된 개념으로, 당시 전통적인 통계 모형으로는 해결할 수 없는 SOR 문제와 같은 어려운 문제들을 해결하는 데 사용되었으나, 구조가 복잡하고 처리 시간이 길기 때문에 당시의 컴퓨팅 자원으로는 효율성이 오히려 떨어지는 연산 방식이었다. 그러나 최근 프로세서 성능의 비약적인 발전으로 기존에 인공신경망 처리 방법이 안고 있던 문제들이 하드웨어적으로 또한 소프트웨어적으로 해결되면서 인공신경망 기법 자체가 구현 가능한 방법이 되자, 기계학습에서 유용한 방법으로 연구자들이 널리 사용하기 시작했다. 최근 그래픽 프로세서(graphic processing unit)를 기반으로 하는 평행처리(parellel processing) 및 심층학습 프레임워크(deep learning framework)가 무료로 보급되면서, 심층학습은 기계학습 내에서

도 특별히 주목받으며 널리 쓰이게 되었다.

이에 따라 인공신경망에 대한 명칭도 시대의 흐름에 맞추어 바뀌고 있는데, 인공신경망이 처음 고안된 1960년대에는 주로 인공신경망 계산법(artificial neural networks computation) 혹은 신경망 계산법(neural computation)으로 불렸다. 하지만 최근에는 인공신경망이 더욱 생체신경망의 구조에 근접해 가고 구조적 복잡도도 증가하면서 심층 신경망(deep neural networks) 혹은 심층학습(deep learning)으로 불린다.

심층학습은 최근 거의 모든 분야에서 각광받고 있으나, 가장 활발하게 사용되는 분야는 컴퓨터 시야(computer vision), 음성 인식(voice or speech recognition), 사진 인식(minage recognition), 자연어 처리(natural language processing) 등이다. 전통적인 통계적 분석 방법에서는 사진이나 영상, 음성, 혹은 자연어를 컴퓨터가 처리할 때 새로운 데이터가 기존의 분류에 정확하게 일치하지 않으면 컴퓨터가 인식하거나 처리할 때 정확성이 매우 떨어졌다. 따라서 연구자들은 새로운 사진이나 음성, 자연어가 주어졌을 때 컴퓨터가 학습을 통해 기존의 데이터들과의 유사성을 스스로 탐색하여 유형을 분류하고 분석을 수행하는 방식을 찾으려 했다. 심층학습은 이를 위한 가장 효율적인 방법은 아닐 수 있으나 현재의 컴퓨팅 자원과 가용한 방법론을 고려했을 때 효과적인 방법론 중 하나라고 할 수 있다.

최근에는 사진이나 영상, 자연어 등의 비정형 자료만을 대상으로 하지 않고 심층학습을 경제 및 경영 분야에서 도입하려는 시도도 이루어지고 있다. 하지만 심층학습은 학습된 결과물을 시스템이 도출하기는 하지만 왜 이러한 결과를 내놓았는지에 대해서는 설명을 제공하지 않는다는 치명적인 약점 때문에 널리 쓰이지는 않고 있다. 하지만 이미 자연어나 사진, 음성 등을 경제학적 분석과 결합하거나, 혹은 해석 불가능성을 완화하기 위한 라쏘 및 릿지 등의 여러 시도들이 이루어지고 있기 때문에 멀지 않은 미래에는 사회과학 및 인문과학에서도 심층학습이 널리 쓰일 가능성을 배제할 수 없다.

2. 자연어 처리를 활용한 선행 연구

본 소절에서는 기계학습 및 심층학습 혹은 자연어 처리를 활용하여 경제 현상을 수치적으로 예측(numerical value prediction)하거나 활용한 사례를 살펴보고 노동시장 예측 모형을 설계하는 데 있어 시사점을 찾고자 한다.

현재 기계학습을 가장 활발하게 활용하는 분야는 금융이다. 주가는 시시각각 변화하기 때문에 주가의 움직임을 예측하거나 분석하기 위해서는 짧은 주기로 양산되는 많은 양의 빅데이터를 처리하거나, 혹은 일반적인 통계 자료 외에 각종 사회관계망(SNS)이나 뉴스 등 자연어 형태의 자료를 분석해야만 한다. 또한 회계 분야에서는 단순한 숫자의 크기나 배열만 분석해야 하는 것이 아니라, 회계 장부에 포함된 자연어들도 분석 과정에서 포함되어야 함은 물론 멀리 떨어진 표와 셀 간의 관계도 중요하기 때문에 동떨어져 보이는 숫자들 간의 관계까지 모두 아우를 수 있는 심층학습이 필수적이라 하겠다. 대형 투자회사의 정보를 이용할 수 있는, 연봉이 고액인 펀드 매니저들의 수익률이 5년 평균 기준으로 주식시장의 평균 수익률을 상회하는 경우가 드물다는 연구 결과(Soe, 2014)가 나오면서, 심층학습을 통해 금융시장에서 투자와 관련된 의사결정을 도와주는 프로그램에 대한 수요가 증가하고 있다.

Vargas et al.(2018)은 주가를 예측하여 투자에 대한 의사결정을 하는 데 자연어 처리 기법을 활용하였다. 이들은 파이낸셜 뉴스의 기사와 기술적 지표들을 활용하여 자연어를 심층학습으로 분석한 후 주가의 일일 변동성을 예측했는데, 이들은 자연어를 포함한 예측 모형이 그렇지 않은 모형과 비교하여 더 나은 결과를 내는지 비교 분석하였다. 이들의 연구 결과에 따르면 파이낸셜 뉴스 기사를 포함한 의사결정 모형은 투자 수익의 안정성은 제고했으나, 수익성 자체를 높이지는 않는 것으로 나타났다.

Cerchiello et al.(2018)은 앞선 Vargas et al.(2018)과 유사하게 파이낸셜 뉴스의 기사와 각종 수치화된 통계 자료를 활용하여 은행의 안정성을 평가하는 모형을 개발하였다. 이들은 파이낸셜 뉴스의 기사를 Doc2Vec[2] 알고리즘으로 벡터화(vectorization)한 후 이를 은행의 재정 상황에 대한

수치 자료와 결합시켰다. 연구 결과, 은행의 재정 상태를 보다 정확하게 파악하는 데 있어서는 단순한 재무제표 및 은행에 대한 각종 수치적 자료만이 아니라 파이낸셜 뉴스를 포함한 모형이 더 나은 성과를 보였다. 앞선 Vargas et al.(2018)과 비교하면, Vargas et al.(2018)은 자연어까지 포함하여 단기 변동성이 큰 수치를 빠르게 예측하는 모형을 만들고자 하였고, Cerchiello et al.(2018)은 발표 주기가 상대적으로 긴 수치에 대한 예측 모형을 개발하였다. 두 연구 결과의 차이는, 자연어 처리가 적어도 단기간에 변동성이 큰 수치를 예측하는 데에는 적합하지 않으나, 중장기적인 stock에 대해서는 상대적으로 예측 효율성 제고(提高)에 도움이 될 수 있음을 시사한다.

경제성장률에 대해서 자연어 처리와 심층 분석을 결합한 연구도 있다. Sheehan et al.(2019)는 위키피디아 기사를 분석하여 개발도상국의 경제 개발 상황을 예측하는 모형 개발을 시도했다. 개발도상국이나 저소득 국가에서 나타나는 행정력의 미비로 인한 행정 통계의 부정확성 및 발간되는 행정 통계의 절대적 숫자 부족 등의 문제를 해결하고 보다 정확하게 저소득 국가의 경제 상황을 알아보기 위해 저자들은 자연어를 분석에 포함시키는 방안을 시도하였다. 이들은 집단의 재산 상황 및 교육 성취도 등 몇 가지 경제 관련 지표들을 위키피디아 기사를 통해 예측했다. 이들이 위키피디아 기사를 대상으로 삼은 것은 저소득 국가에 대해 편향이 적으며 신뢰할 수 있는 자연어 자료를 얻기 위해서이다. 이들 연구에서는 자연어를 포함하지 않은 모형과 비교하여 자연어를 처리한 모형에서 예측력이 더 높은 결과가 나왔다.

한국어를 직접 사용하여 경제 현상을 예측하거나 설명하려는 시도는 아직 많지 않은 실정이다. 일부 북한 관련 연구(김수현・손욱, 2020)에서 자연어 처리 기법이 시도되었으나, 아직은 수치화된 통계 자료를 활용하여 심층학습을 통해 모형을 개발하는 연구는 소수(少數)에 머물러 있다.

---

2) Doc2Vec은 자연어 처리 기법 중 하나로, 문자로 작성된 자연어를 컴퓨터가 분석 가능하도록 수치적으로 변환시키는 여러 기법 중 하나이다. Doc2Vec의 원리는 Le, Q. V. & T. Mikolov(2014), “Distributed Representations of Sentences and Documents,” 32에 보다 상세하게 설명되어 있다.

Soohyon Kim(2020)은 앙상블 러닝과 베이지안 러닝 기법으로 심층학습을 통해 거시경제 변수인 월간 통관기준 수출과 환율을 예측하는 모형을 개발하였다. 그 결과, 심층학습 모형은 두 수치 모두 더 나은 예측력을 보여줬음은 물론, 예측치의 오차 범위도 보다 정확하게 제시하였음을 보였다.

정한웅(2016)은 심층학습 알고리즘을 활용하여 1999년부터 2015년 사이에 국내 유가증권시장에 상장된 비금융 기업의 부도를 예측하는 모형을 개발하였다. 연구 결과에 따르면 심층학습 알고리즘은 기존의 부도 예측 모형들과 비교하여 더 나은 예측력을 보여주었다.

이상에서 기존의 자연어 처리 및 심층학습을 이용하여 경제 현상을 분석한 연구들을 살펴본 결과, 자연어 처리는 단기간에 빠르게 변화하는 수치를 예측하는 데에는 부적합할 가능성이 높다. 하지만 수치화하기 어려운 여러 사회경제적 지수(socioeconomic indicators)나 충분한 기간을 두고 집계되는 저량(stock variable)에 대해서는 충분히 나은 예측력을 가질 수 있음을 시사하고 있다. 또한 자연어 처리를 하지 않더라도, 심층학습은 시스템이 스스로 오차를 개선함으로써 모형을 개선하여 전통적인 경제 모형과 비교하여 더 나은 예측력을 보였음을 확인할 수 있다. 다만, 심층학습을 통한 예측이 자칫 과적합하는 문제로 인해 설명력이 높은 것일 수도 있어, 충분한 기간을 두고 수치를 예측케 하여 장기에 걸쳐서도 심층학습 모형의 예측력이 높은지, 그리고 급변하는 상황이 발생했을 때에도 좋은 예측력을 보여주는지에 대해서 검토해 가며 예측에서의 효율성을 검토할 필요가 있다 하겠다.

또한 지금까지 자연어 처리를 통해 경제 현상을 예측하거나 분석한 연구 다수가 영어를 대상으로 하였으나, Doc2Vec 및 여러 자연어 처리 기법의 원리상 자연어 처리가 특정 언어에 보다 적합한 방향으로 설계되어 있는 것은 아니기에 한국어에 대해서도 자연어 처리 기법을 적용시켜 그 결과를 보는 것이 의미가 있다 하겠다. 특히, 이를 통해 향후 한국어에 대해서 자연어 처리 성능을 향상시키는 방법이나 영어와 비교한 분석에서의 정확성 및 수월성 등도 이후의 자연어 처리 연구에서 참고할 수 있을 것으로 보인다.

3. 기계학습을 이용한 자연어 처리

최근 사용하는 자연어 처리 기법은 다수가 기계학습, 그중에서도 특히 심층학습을 기반으로 하고 있다. 심층학습은 많은 컴퓨팅 자원(computing resource)를 요구하지만, 기존에 널리 사용하던 통계적 모형을 기반으로 한 자연어 처리 방법들보다 더 정확성이 높은 예측 및 분석 결과를 보여주고 있다. 대표적인 자연어 처리 기법 적용의 예로 자동 번역 시스템이나 자동 문서 요약 시스템, 혹은 Q&A 시스템을 꼽을 수 있으며, 최근 확산되고 있는 자동 안내 기능 및 자동 응답 기능 역시 이러한 자연어 처리의 한 예라 할 수 있다.

자연어 처리가 효과적으로 이루어지기 위해서는 자연어로 구성된 자료를 정리하는 전처리 과정(pre-processing)이 중요하다. 잘 정리되지 않은 데이터가 입력되면 기계학습은 필연적으로 분석 기간이 오래 걸리거나, 분석에서 정확성이 떨어지거나, 혹은 둘 모두가 발생한다. 따라서 자연어 처리에 있어서 전처리는 분석의 효율성과 결과의 정확성 제고를 위해 필수적이다.

데이터 전처리 과정에서 주로 이루어지는 것은 데이터에는 포함되어 있지만 기계학습과는 무관한 불용 단어(stopwords), 문장 기호(punctuations), 특수문자(special characters)를 제거하는 것이다. 이때 모든 문자 기호나 특수문자를 제거하는 것은 아니며, 프로그램이 중요하게 인식할 수 있는 것은 남겨놓아야 한다. 예를 들어, 문장 단위로 자연어를 분석하는 데 있어 문장의 종료를 알리는 마침표는 지워서는 안 되며 반드시 필요하다 할 수 있다. 하지만 문장의 종결을 선언했음에도 반복적으로 종결 기호가 등장하거나, 충분한 의미 전달이 이루어졌음에도 각종 특수 입력 기호를 삽입한 경우에는 이들을 삭제해야 한다.

한편 데이터 전처리 과정에서 이후 자연어 처리를 위해 반드시 필요한 과정이 형태소 분석(tokenization)이다. 시스템이 자연어를 처리하기 위해서는 입력된 언어를 최소 처리 단위(형태소, token)로 분할해야 한다. 만일 여러 문장들로 구성된 텍스트를 다루는 경우에는 형태소 구분 이전에 문장 단위 구분 작업을 수행한 후 문장 단위로 형태소를 구분한다.

전처리를 마친 자연어 텍스트는 기계학습 알고리즘에 의해서 처리된다. 기계학습에 의해서 자연어가 처리되는 과정은 크게 인덱싱(indexing), 수치 변환(word embedding), 패딩(padding), 그리고 텐서(tensor) 변환을 거친다.

자연어는 그 자체로 시스템이나 프로그램이 읽어서 처리할 수 없다. 따라서 기계가 읽을 수 있는 언어로 자연어를 변환해야 하는데, 입력된 데이터를 기계어로 변환하기 위한 과정 중 하나가 인덱싱이다. 인덱싱 과정에서 모든 단어들은 고유의 인덱스 값으로 대체되어 처리된다. 다시 말해서 개별 단어나 표현마다 1:1의 고유한 값을 부여해 이후 수치로 변환한다. 이때 0번 인덱스는 특수 문자인 '<pad>'에 할당하는데, 이는 이후 설명할 패딩 과정을 위해서 필요하다. 따라서 대부분의 인덱스 과정에서 0번으로 할당되는 표현은 없다고 보아도 좋다.

인덱싱을 마치면 시스템은 자연어를 필요한 구성 요소별로 분해해서 각각의 구성 요소를 기계가 읽을 수 있는 수치(numerical value)로 변환하는데 이 과정을 수치 변환이라고 한다. 예를 들어, '기계학습을 통해 노동시장을 분석한다'는 문장을 컴퓨터가 처리한다면, 하나의 문장을 ['기계', '학습', '을', '통해', '노동시장', '을', '분석', '한다']라는 8개의 구성 요소로 분할하여 각각의 구성 요소별로 고유의 벡터값을 배정한다. 이렇게 개별 구성 요소별로 수치화된 벡터값을 부여하는 과정을 수치 변환이라 한다. 입력된 자연어의 개별 단어를 고유한 벡터로 변환하는 수치 변환 과정에서, 각각의 고유 벡터는 특정한 통계적 가정하에 임의로 배정될 수도 있으나, 일반적으로 위키백과 등과 같은 대량의 자연어 조합(corpus)을 가지고 미리 기계학습한 벡터(pre-trained word embeddings)를 사용하는 것이 수치 예측 등과 같은 자연어 처리를 더 정확하게 수행하는 데 도움이 된다는 것이 학계의 대체적 의견이다.

이론적으로는 수치 변환 이후에 개별 표현이나 문장마다 길이가 다를 수밖에 없으나, 심층학습을 위해서는 이들이 모두 동일한 길이의 벡터를 가져야 한다. 따라서 길이가 가장 긴 표현이나 문장을 기준으로 삼아서, 그보다 길이가 짧은 문장이나 표현은 벡터의 길이를 늘려줘야 한다. 앞서 인덱스에서 0번은 '<pad>'에 할당된다고 했는데, 이 '<pad>'를 통해서 짧

은 표현이나 문장들의 벡터 길이를 모두 동일하게 맞춘다. 이러한 과정을 기계학습에서는 패딩이라 한다. 패딩을 마치고 나면 모든 벡터들은 동일한 길이를 가지며 시스템이 분석할 수 있는 준비를 거의 마치게 된다.

기계학습을 위한 수치화의 마지막 단계는 텐서 변환이다. 원래 텐서는 일반화된 n차원 행렬(generalized n-dimensional matrix)을 의미하지만, 기계학습에서는 3차원 이상의 숫자 배열을 텐서라고 부른다. 텐서로 변환된 자연어들은 심층학습 프레임워크의 자료 처리 기본 단위인 텐서로 입력되어 프로그램이 읽고 처리한다.

텐서화된 자연어들은 심층학습 예측 모형을 통해서 분석된다. 심층학습 내에서는 자연어를 처리하는 여러 가지 모형이 존재하지만, 최근의 자연어 처리 연구에서는 크게 순환 신경망(recurrent neural network : RNN) 기법과 합성곱 신경망(convolutional neural network : CNN) 기법이 주로 사용된다. 자연어는 문장의 순서가 중요하며, 순서가 바뀌면 의미가 성립하지 않거나 혹은 의미가 바뀌는 경우가 많다. 또한 문장이나 표현의 길이가 가변적이고, 특정 문장이나 표현은 유달리 긴 경우도 있다. 순차적인 자료의 해석과 가변적인 길이의 자료를 처리하는 데 있어서는 LSTM(Long Short-Term Memory)나 GRU(Gated Recurrent Unit)와 같은 순환 신경망이 합성곱 신경망보다 적합하다는 것이 학계의 중론이다.

합성곱 신경망은 자연어의 이해와 처리를 위해 필수적인 문맥의 장기의존성(long-term dependency of the context)을 파악하기 어렵기 때문에 활발하게 사용되고 있지는 않다. 또한 합성곱 신경망은 순환 신경망에 비해 상대적으로 보다 큰 용량의 메모리를 요구한다는 물리적인 약점도 존재한다. 하지만 합성곱 신경망은 순환 신경망에 비해 연산이 신속하고 효율적이며, 최근 대용량의 메모리를 확보하는 것이 이전과 비교하여 수월해진 점 때문에 점차 활용 빈도가 증가하고 있다.

하지만 모든 순환 신경망 기법이 자연어 처리에 좋은 것은 아니다. 특히 긴 텍스트나 문헌을 탐색할 때는 단순 순환 신경망(simple RNN)을 사용하는 것이 적합하지 않을 수 있다. 단순 순환 신경망 모형은 연산이 단순하고 효율적이며 과거의 이벤트를 현재의 예측에 반영한다는 장점은 있지만, 과거의 이벤트를 반영하는 고려 가능 기간이 상대적으로 짧다.

반면 LSTM은 계산에 필요한 매개 변수의 숫자가 더 많기 때문에 더 높은 계산 능력을 요구함에도 오래전에 발생한 이벤트를 현재의 예측 모형에 사용할 수 있다는 장점이 있다. 다수의 심층학습 사례에서 단순 순환 신경망으로 구성한 모형을 LSTM 모형으로 대체하는 것만으로도 예측의 정확성이 높아진다고 알려져 있다.

제2절 데이터

1. 정형 데이터

빅데이터 분야에서는 데이터의 유형을 크게 정형 데이터(structured data)와 비정형 데이터(unstructured data)로 나눈다. 정형 데이터란 수치화된 결과물을 제공하는 통계 자료를 의미한다. 예를 들어, 분기별 경제성장률 수치는 거시경제의 변동성을 보여주는 정형 데이터라 할 수 있다. 비정형 데이터는 수치화되지 않아 컴퓨터가 바로 읽고 처리할 수 없는 데이터를 의미한다. 대표적인 비정형 데이터가 문자, 사진, 영상, 음성 등이다. 시스템이 비정형 데이터를 처리하기 위해서는 일단 수치화 작업을 통해 정형 데이터와 유사한 형태로 만든 후 처리한다. 본 소절에서는 분석에 사용한 정형 데이터는 무엇인지 살펴보고 소개할 것이다.

정형 데이터는 크게 두 분류로 나눌 수 있다. 첫째는 노동시장 관련 통계로, <표 2-1>은 분석에서 사용한 통계 목록을 보여주고 있다. 이 중 경제활동참가율, 실업률, 고용률은 실제 모형이 예측하고자 하는 변수들이다. 각 변수들은 성별의 경우 전체와 남성 및 여성에 대해서, 연령별 변수들은 제공하는 모든 연령대별 변수 각각에 대해서 통계치를 사용하였다. 경제활동인구조사에서 제공하는 연령 유형은 15세 이상 전체, 15~19세, 20~29세, 30~39세, 40~49세, 50~59세, 60세 이상, 15~64세, 15~24세, 15~29세이다. 교육정도별 경제활동인구 총괄에서 제공하는 교육수준 유형은 전체, 초졸 이하, 중졸, 고졸, 대졸 이상, 전문대졸, 대학교졸 이상 이

렇게 7가지이다.

기계학습은 경제활동인구 및 각종 노동지표의 성별 변화, 연령별 추세, 교육정도별 트렌드 등을 분석하고, 아울러 당기에 각 통계가 실제 예측하고자 하는 통계와 어떠한 관계가 있는지를 파악할 것이다. 그래서 현재까지 파악된 연관성을 바탕으로 예측하고자 하는 시기의 노동통계를 제공할 것이다.

이외에 분석에서 사용하는 통계는 인구 이동, 경기 및 경제에 관한 변수, 생산 및 투자 관련 지수, 재고 지수, 철강 등 몇몇 품목의 생산량, 공종별 혹은 동수별 건설수주액과 수주 건수, 산업별 품목별 수출입 물량 및 수출입 물가지수, 품목별 소비자물가지수와 생산자물가지수, 신용카드 사용액, 각종 지급 결제 및 통화금융 관련 통계들, 그리고 주식시장에 대한 지표들을 포함하고 있다. 이들 지표는 모두 2001년 1월 이후 2020년 10월까지 모두 월별 단위로 측정된 지표이다. 일별 지표로는 2001년 1월 1일부터 2020년 10월 31일까지의 일자별 시중 각종 금리 지표 및 주가지수를 사용하였다. 사용한 변수들의 상세한 목록은 [부록]에 첨부하였다.

정형 데이터 다수는 실물 경제의 움직임이나 사회 내에서 발생하는 여러 변화의 방향 및 크기를 실질적으로 측정한 결과를 나타낸다. 대표적으로 광업 및 제조업에서의 산업별 내수출하지수나 수출출하지수, 소비재

〈표 2-1〉 노동시장 관련 통계

변수명	추출 대상	제공 기간	출처
성별 경제활동인구 총괄	경제활동인구, 경제활동참가율, 실업률, 고용률	2001. 1 ~ 2020. 10	통계청, 「경제활동인구조사」
연령별 경제활동인구 총괄			
교육정도별 경제활동인구 총괄			

자료 : 통계청, 「경제활동인구조사」.

나 자본재, 그리고 중간재별 제조업 생산지수, 월별 철강 생산량 등이 이러한 변수들의 대표적인 예이다. 또한 임금 및 물가 관련 통계는 생산과 소비에 대한 물가 정보를 제공함으로써 간접적으로 거시경제 환경 및 생산 시장 상황을 알려줌과 동시에, 물가 변동에 따른 노동 공급자들의 노동시장 참여 의사의 변화를 간접적으로 모형에서 반영해 줄 것이다. 결제 시스템에서의 결제 건수와 액수 등에 대한 통계는 거래가 얼마나 많이 일어나고 있으며 대외적으로는 얼마나 많은 거래가 발생했는지를 보여주는 통계로, 만일 거래가 활발하다면 경제 활동이 많아짐에 따라 경제활동참가율이나 고용률은 올라가고 실업률은 낮아질 가능성이 높다. 반대로 국내 거래량도 적어지고 수출도 감소한다면 경제활동참가율이 낮아지고 고용률이 떨어지며, 실업률은 높아질 것이라 예측할 수 있다. 에너지 관련 통계는 국내 산업 생산 및 경제활성화의 정도를 반영한다. 에너지 관련 생산 및 소비 지표는 계절성을 띠는 관계로 계절성에 따라 노동 지표의 변화를 예측할 때 단순히 계절조정을 하는 것과 비교하여 보다 예측력을 높이고, 에너지 사용량의 변화를 통해 국내 경제활동의 정도를 간접적으로 측정할 수 있다.

그러나 모든 정형 데이터가 실물 경제만을 반영하는 역할을 하는 것은 아니며, 일부 자료는 사람들의 경기 상황에 대한 인식이나 전망을 포함하고 있다. 이러한 변수들의 예로 경기종합지수, 경제심리지수, 그리고 기업경기조사 전국전망을 들 수 있다. 경제 및 경기 관련 변수들은 기업이 현재의 경기 상황을 어떻게 바라보고 있으며 앞으로 경기가 어떠할 것으로 예상하는지를 담고 있으므로 기업의 고용 및 해고와 일정한 상관관계가 있을 수 있다. 즉, 경제나 경기의 상황이 좋지 않으면 현재의 고용 관련 지표가 좋지 않을 것이며, 전망 지표가 좋다면 적어도 단기에서는 노동시장 지표들이 현재보다 개선될 것이라 예측할 수 있다.

기업경기조사 전국실적은 산업별로 업황, 매출, 채산성, 자금사정, 인력사정에서의 실적을 제공한다. 기업경기조사 상의 산업 유형은 전산업, 제조업, 비제조업 세 가지이다. 기업경기조사 전국전망은 동일한 산업과 항목에 대해서 전망치를 물어본 것인데, 전망 지수에 대해서는 자료 특성상 데이터를 한 개월 앞으로 매칭하였다. 기업경기조사 실적과 전망은 동일

한 기간에 동일한 항목을 물어보았으나, 실적은 당월 실적이며 전망은 익월에 대한 값이므로 두 통계는 실적이 전망보다 1개월 앞선다. 즉, 2010년 7월에 조사한 값에 대해서는 실적 조사는 7월 값으로 간주되나 전망 조사는 8월 값으로 간주된다. 전망 지수는 조사 시점 기준 익월에 대한 전망이지만, 전망 자체가 해당 월의 제반 상황을 고려하여 이루어진 것이므로 실제로는 해당 월의 여러 특성을 대변한다고 볼 수 있다. 따라서 익월에 대한 전망치로 제시된 통계를 당월로 1개월 당겨서 변수값으로 사용하였다.

경제성장률 등을 포함한 성장 관련 거시경제 지표와 일부 분기 및 연단위 지표는 분석에서 사용하지 않았다. 현재 예측하고자 하는 변수들은 모두 월단위 변수인 데 반해 분기 단위 자료를 사용하는 경우에는, 분기 단위 자료의 업데이트 주기가 느려 불가피하게 아직 발표되지 않은 2020년 3/4분기나 하반기, 혹은 2020년 연단위 지표를 사용할 수 없다. 이 경우 예측을 위해 가장 중요한 최근의 관측치들이 결측치(missing value)로 표시된다.

물론 기계학습에서는 결측치를 처리할 수 있는 몇 가지 방법이 있다. 하지만 학습을 위해 사용되는 과거 값이 결측치인 경우에는 예측력에 큰 영향을 미치지 않을 수 있으나, 최근의 통계치가 결측된 경우에는 최근 값에 가중치를 크게 부여하는 기계학습의 특성상 예측에 큰 영향을 주기 때문에 결측치가 없는 월별 데이터만 사용하고 연도별, 반기별 및 분기별 자료는 사용하지 않았다.

예측하는 종속변수 및 주요 변수값들이 월별 통계량인 반면, 일부 변수들이 일별 통계이기 때문에 분석 과정에서 시스템을 통해 이들을 매칭할 필요가 있다. 이때는 크게 두 가지 방법이 있다. 하나는 월별 데이터를 일별 데이터에 연계하여 월별 값들이 해당 월에는 매일 할당되는 것으로 하여 일별 자료에 월별 자료를 매칭시키는 것이 하나이다. 다른 방법은 일별 데이터를 월별로 압축한 후 월별 자료와 매칭하는 것이다. 여기서는 예측 변수가 월 단위 변수이므로 일별 자료의 경우에는 월 단위로 변환한 후 이를 다른 월별 자료와 함께 활용하여 예측토록 하였다. 보다 상세한 방법은 다음 소절에서 비정형 데이터를 어떻게 처리했는지 설명하면서

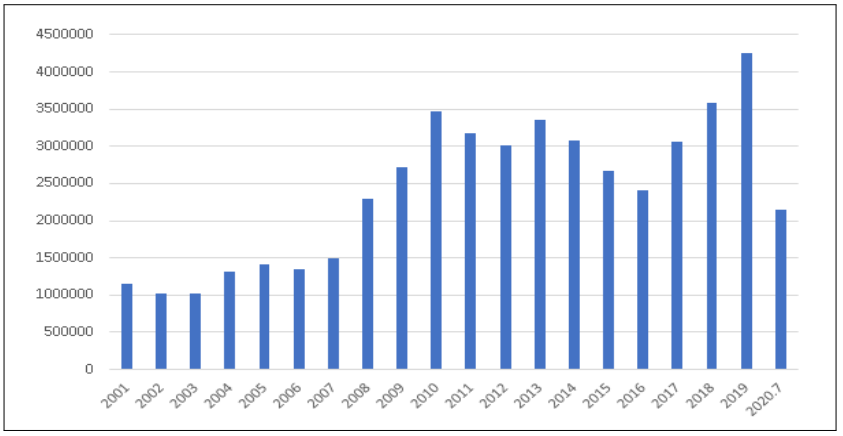
소개할 것이다.

2. 비정형 데이터

본 연구에서는 정형 데이터들뿐만 아니라 비정형 데이터도 분석에서 활용하려 한다. 분석에서 사용하는 비정형 데이터는 2001년 1월 1일 이후 생산된 모든 신문 기사 제목이다. 정형 데이터는 실물 경제의 변화와 아울러 제한적으로 경제 주체들이 경기 전망을 어떻게 하고 있는지와 현재 경제에서 관심사가 무엇인지를 간접적으로 혹은 단기적으로는 보여주지만, 이에 대한 직접적인 자료는 제공하지 않는다. 신문 기사는 일반 경제 주체들의 여론 동향 및 관심사항에 대해서 보다 직접적인 정보를 제공할 것이다.

실제 분석에서 사용한 비정형 데이터는 2001년 1월 1일부터 2020년 10월 31일까지 238개월에 걸친 신문 기사의 제목을 수집한 것이다. 신문 기사 제목의 전체 분량은 총 48,353,727건이며, 이들의 연도별 분포는 [그림 2-1]과 같다. 오래된 신문 기사의 경우 기사의 삭제 및 수정으로 인하여 잔존한 전수를 수집하였음에도 2010년대와 비교하여 적음을 알 수 있다. 이러한 경향은 특히 2008년 이전에 더욱 뚜렷한 것을 확인할 수 있다.

〔그림 2-1〕 연도별 수집 기사 수



자료 : 저자 작성.

앞서 설명하였듯이 비정형 데이터는 컴퓨터에 의해서 인식되고 분석되기 위해서 전처리 과정을 거쳐야 한다. 전처리 과정이 적으면 적을수록 본래의 원자료에 가까운 데이터가 활용되지만 입력되는 자료의 품질이 떨어져 분석의 정확성이 저해될 수 있다. 하지만 임의의 많은 전처리는 데이터가 담고 있는 정보량 자체를 줄일 수 있을 뿐만 아니라 본 연구의 목적인 기계학습에 의한 시스템의 비정형 데이터 분석이라는 본래의 목적을 저해할 가능성이 있으므로, 연구자의 자의적 개입을 줄이기 위해 최소한의 전처리만을 수행하였다.

전처리 과정에서 가장 먼저 수행한 것은 자료 재분류 및 선별이었다. 데이터 검토 과정에서 스포츠나 연예 신문사의 경우에는 데이터의 분석과 무관한 사진 자료나 개인 대소사(부고/인사/혼례)와 같은 자료들이 많이 섞여 있었다. 사진의 경우 딥러닝에 의해 학습 후 처리할 수 있으나, 본 연구에서는 한국어 텍스트 처리에 초점을 맞추었기 때문에 사진 등에 대해서는 별도의 딥러닝을 하지 않을 예정이다. 따라서 이러한 기사들을 제외시키기 위해서 스포츠 및 연예지와 지방 신문사를 제거하였다.

신문사를 기준에 의해서 제거한 후에도 중앙 일간지에서 여전히 사진 자료가 포함된 신문 기사 제목이 발견되어 문서화된 제목을 제외한 나머지 자료는 필터 처리하여 제외하였다. 이후 보다 고용량의 컴퓨팅 자원을 사용할 수 있게 된다면 이들 사진 자료까지 처리하는 것도 가능할 것이다.

사진 자료를 삭제한 후 남은 신문 기사 제목에 대해서 한글, 한자, 영어, 숫자 및 단위를 표시하는 문자(% 등)를 제외한 문장 기호 및 부호는 분석과 무관하므로 제외하였다. 문장이나 문단 단위로 비정형 데이터를 분석하는 경우에는 최소 분석 단위의 식별을 위해서 문장 기호를 삭제해서는 안 된다. 하지만 본 연구에서는 신문 기사의 제목만을 별도 수집하여 제목마다 식별 가능한 형태로 가공하였기에 문장 기호는 필요치 않았다. 대신 자연어 처리를 위한 최소 처리 단위인 형태소(token)는 식별할 필요가 있다. 예를 들어, '경제 안 좋다는데 이상하네 2100 뚫은 코스피 3가지 이유'와 같은 문장은 ['경제', '안', '좋다는데', '이상하네', '2100', '뚫은', '코스피', '3', '가지', '이유']의 10개의 형태소로 분할된다. 일반적인 텍

스트 분석에서는 형태소를 분석하기 이전에 보다 큰 단위인 문장 분석을 수행한다. 하지만 본 연구에서는 문장 단위의 구분이 필요치 않으므로 이를 생략하고 바로 형태소 분석을 실시하였다. 형태소 구분은 Konlpy 패키지 안에 포함되어 있는 유호현 박사에 의해서 개발된 오픈 소스(open source) 형태소구분기인 Okt Class를 사용하였다.

짧은 신문 기사 제목의 경우 다수가 의미 있는 내용을 담고 있지 못하는 경우가 많았다. 이 경우 이들을 포함시키는 것은 노이즈(noise)로 잡혀 모형의 정확성을 떨어뜨리는 경우가 많다. 따라서 상술한 과정을 거친 후에 형태소의 길이가 세 개 미만인 신문 기사 제목은 모두 삭제하였다.

한편 신문 기사 제목을 사용할 때 포함된 항목은 일자로 표시된 신문 기사의 발행 시기와 신문 기사 제목, 그리고 발간 신문사이다. 발간 신문사는 신문사별 논조 및 신문사에 따른 관심 사안과 집중도의 차이를 식별해서 비정형 데이터의 분석 정확성을 높이기 위해 수집하였다. 반면 정치, 경제, 사회, 문화 등 신문 기사의 분류는 포함시키지 않았다. 이는 신문 발행사마다 분류 체계가 다르며, 몇몇 기사의 경우 한 개 이상의 분류를 가진 것들도 있었다. 이 경우 해당 기사를 단일 분류로 만든다면 임의성이 개입되며, 2개 이상의 분류를 허용할 경우, 해당 신문 기사가 중복으로 분석에 포함될 수 있다. 따라서 두 가지 가능성을 배제하기 위하여 신문 기사의 분류 항목은 분석에서 배제하고 단일 기사 각각이 1개만 반영되도록 했다.

제3절 예측 모형 설계

1. 프로그램의 구성

본 연구에서 예측하려는 노동시장 변수는 월별 경제활동참가율, 월별 고용률, 월별 실업률 세 개이다. 이들을 예측하기 위해서 사용할 모형은 세 가지이다. 하나는 노동시장 변수들만을 가지고 학습한 후 이를 바탕으

로 최근 5개월간의 노동시장 변수를 예측하는 모형이다. 둘째는 노동시장 변수를 포함한 모든 정형 데이터를 이용하여 학습한 후 노동시장 변수를 예측하는 모형이며, 마지막은 모든 정형 데이터에 비정형 데이터까지 포함하여 노동시장 변수를 예측하는 모형이다.

노동시장 변수를 예측하기 위한 모형은 파이썬(python)에 설계하여 구동하였다. 시스템은 총 4개의 파이썬 프로그램과 변수 설정을 위한 파일 1개로 구성되어 있다.

첫 번째 프로그램은 데이터를 읽기 위한 프로그램이다. 모든 정형 자료는 엑셀 파일로 구축하였는데, 엑셀 파일로 구성된 원시 데이터(raw data)를 파일로부터 읽어내 데이터 전처리 과정을 거치고 인덱싱을 한 후, 처리된 데이터를 심층학습 프레임워크에서 텐서플로우(tensor flow) 프로세싱의 기본 구성 단위인 텐서로 변환시키는 역할을 한다.

두 번째 프로그램은 심층학습 모형을 구축하는 과정을 수행하는 프로그램이다. 심층학습은 여러 개의 층(layer)으로 이루어져 있는데, 각각의 층에 대한 정의와 층을 통해 처리되는 텐서들의 형태 및 텐서를 처리하는 유닛(unit)을 지정하는 역할을 수행한다.

세 번째 프로그램은 전처리된 자료들을 모형을 통해서 학습하는 학습 프로그램이다. 학습 프로그램은 개별 학습 과정(epoch)이 종료될 때마다 모형을 통해 얻은 값과 실제 값과의 차이를 사전에 정의한 손실함수를 이용해서 계산한다. 이 학습 과정은 주어진 자료를 가지고 예측하고자 하는 변수들을 잘 설명하거나 예측하기 위해 모형을 개선하는 과정이다. 여기서 예측과 설명이 혼재되어 있는 이유는 학습 프로그램은 학습 결과가 얻어지면 학습 결과를 실제 종속변수의 값과 대조한다. 그래서 차이가 작은 모형을 계속해서 찾아나가는 과정인데, 이 과정은 학습 데이터용으로 주어진 예측 변수들의 과거치를 가지고 모형을 개선하는 과정이기 때문에, 정확하게는 주어진 종속변수들을 설명하는 모형을 찾아가는 과정이다. 그러나 모형을 설계한 후 설계한 모형을 토대로 종속변수의 값을 과정마다 예측하고 예측치와 실제치를 비교해 가며 개선해 나가는 것이다. 따라서 학습 데이터를 부여한 사람의 관점에서는 모형은 설명을 해나가는 것이지만, 프로그램은 매 과정마다 학습 데이터를 통해 예측을 수행하고

있다.

미리 설정한 손실함수는 프로그램의 예측치와 실제값을 비교해 가면서 미리 설계된 훈련 알고리즘(optimizer)에 의해 모형 내 각 층에 존재하는 계수값들을 업데이트하도록 전파한다. 이러한 전파 과정을 back-propagation이라고 하는데, 학습에 의해 얻어진 결과를 바탕으로 모형을 개선할 점을 바로 직전 층으로 전달한 후, 학습 자료로 돌려서 과정을 반복하기 때문에 back-propagation이라고 한다. 이 과정은 모형에서 반복 수행을 통한 모형의 성과 개선이 더 이상 없다고 판단될 때까지 계속된다.

마지막 프로그램은 모형에 사용될 변수를 읽고 실제로 훈련을 시키는 메인 프로그램이다. 이렇게 네 개의 프로그램에 의해 모형이 구성되며, 모형에 사용될 자료들의 경로 및 변수들을 설정하는 읽기 전용 파일로 parameters.ini 파일을 설정하였다.

이렇게 설정된 모형은 2001년 1월 1일부터 2020년 6월 30일까지의 자료를 가지고 학습한 후, 학습된 모형을 가지고 2020년 7월부터 11월까지 노동시장 변수들의 값을 예측하였다. 따라서 학습을 위해 전체 238개의 데이터를 사용하였다. 일반적으로 기계학습에서는 과적합을 방지하기 위해 요구되는 학습 및 훈련 데이터의 크기를 약 20만 개 정도, 혹은 최소한의 경우에도 약 10만 개 정도로 본다. 하지만 경제 통계는 통계의 특성상 월간 자료가 1년에 12개밖에 생산되지 않으므로 훈련을 위해 사용할 수 있는 자료의 크기가 제한적이라는 문제점이 있다. 이러한 점을 감안하고 다음 장에서 모형별로 예측치를 살펴보고 예측력을 비교해 보도록 하겠다.

2. 학습 방법 설계

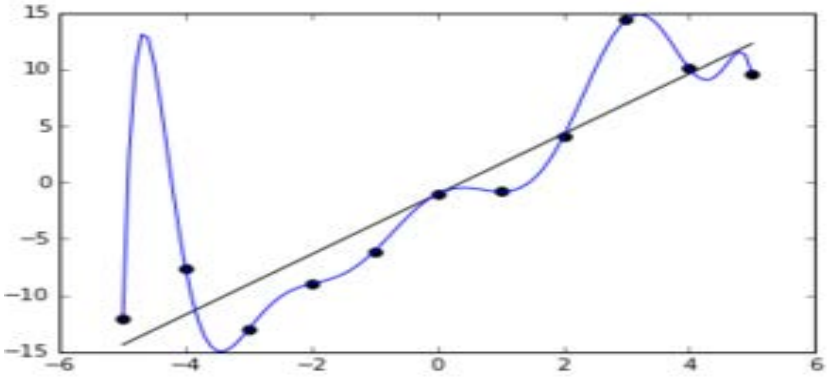
본 연구에서는 손실함수의 값이 최소화가 될 때까지 학습을 반복하도록 하였다. 그러나 이 경우에는 두 가지 문제가 발생할 수 있다. 첫째는 학습 데이터에 맞는 모형을 찾기 위해서 과적합된 모형이 최적화된 모형으로 제시될 가능성이다. 또 다른 문제는 모형이 손실함수의 값을 한없이 줄이기 위해서 계속해서 학습을 반복하여 학습에 많은 시간이 소요될 수 있다. 따라서 학습의 효율성을 제고하기 위해 일반적으로 모형이 학습할

때 drop-out rate과 조기 종료(early stopping)를 설정한다. drop-out rate은 주어진 학습 자료를 이용하여 모형을 탐색한 후 전체 설명변수 중에서 미리 설정된 비율만큼을 임의로 선택하여 삭제한 후 남은 설명변수만을 가지고 모형의 유효성을 검증하는 것이다. 매 회차마다 삭제하는 변수가 달라지며, 이 과정을 최소 일정 횟수 이상만큼 반복하기 때문에 학습 횟수만 충분하다면 모형이 편의(biased)될 가능성은 적다.

drop-out rate을 높이면 유효성을 검증하는 데 있어 사용하는 변수들이 줄어들기 때문에 주어진 학습 함수를 가지고 보다 일반적인 모형(generalization)을 찾으며, 학습 시간도 줄어든다. 하지만 모형의 설명력이나 적합도는 낮아지게 된다. 반대로 drop-out rate을 낮추면 모형의 설명력이나 적합도는 높아지는 대신 학습 시간이 더 길어지며 과적합이 나타날 가능성이 더 높아진다. 이를 쉽게 도식화한 것이 [그림 2-2]에 표시되어 있다.

[그림 2-2]에서 점으로 찍힌 것은 실제값이다. 가장 일반화된 모형은 제시된 점들을 바탕으로 적합선(fitted line)을 그린 일직선이다. 그러나 적합선에서 제시하는 예측치는 실제값과 일치하는 경우가 적다는 문제가 발생한다. 대신 여러 설명변수를 고려하지 않아도 실제값을 가장 직관적이면서도 단순하게 표시한다는 장점이 있다. 이 적합선은 단순하게 x축의 값과 y축의 값 두 가지만을 사용하여 가장 일반적인 모형을 구축한 경우이다.

〔그림 2-2〕 모형의 일반화와 과적합



자료 : 저자 작성.

반면 파란 실선은 실제 점들을 모두 통과하는 실선이다. 이 경우 개별 관측치를 실제로 설명하는 설명력은 높지만, 첫 두 개 값을 연결하는 과정에서 한 번의 점프가 관찰되며, 두 번째 값에서 세 번째 값으로 갈 때 최하점을 거친 후 상승하는 형태로 예측한다. 이는 과도하게 모든 실제값을 설명하기 위해 주어진 자료들을 가지고 적합도를 높이려고 한 결과 나타나는 현상이다. 따라서 파란 실선은 모든 설명변수를 사용하여 과적합된 모형이 실제값을 어떻게 예측하는지를 보여주는 실례(實例)라 할 수 있다.

모형을 통해 다음 값을 예측할 때 과적합된 모형이 반드시 정확한 값을 보여주는 것은 아니다. 제시된 [그림 2-2]에서도 다음 값을 예측할 때 가장 일반화된 모형인 적합선은 최근 값에서 증가한 어떤 값을 제시할 가능성이 높지만, 과적합된 파란 실선은 추세의 우측을 고려할 때 최근 값에서 하락한 어떠한 값을 제시할 가능성이 있다. 따라서 기계학습 모형을 설계할 때는 모형의 적합도를 높이면서도 과적합을 방지하기 위해 동시에 일반성을 높여야 한다. 따라서 drop-out rate은 일반적인 기계학습 모형의 설정을 따라 30%로 하였다. 즉, 모형을 통해 예측을 한 번 수행한 후 앞서 언급했던 back-propagation을 하는데, 이렇게 해서 업데이트된 모형은 전체 설명변수 중에서 임의로 30%가 제거된 남은 70%의 독립변수만으로 다시 모형의 적합성을 검증한다.

조기 종료란 학습을 통한 모형 탐색을 언제 중지할지를 설정하는 기준이다. 기계학습은 예측 모형 혹은 분석 모형을 훈련 데이터에 반복적(iteration)으로 노출시킴으로써 이루어진다. 학습을 반복적으로 수행하면 어느 수준까지는 모형의 일반적인 예측력을 증가시키는 방향으로 최적화가 이루어지지만, 지나치게 오래 반복하면 훈련 데이터에만 최적화되는 과적합의 문제가 발생한다. 따라서 언제 학습을 종료할 것인가, 혹은 어떠한 모형을 최적화된 모형으로 결정하는가에 대한 문제가 제기되는데, 훈련 종료 시점 파악을 위해 모형의 개선이 완료된 후 테스트 데이터를 사용하여 해당 훈련 모델이 얼마나 일반성을 유지하면서 좋은 예측력을 가지고 있는지를 매번 반복하여 점검한다. 점검 결과, n번 이상 연속적으로 테스트 데이터에서 예측력이 증가하지 않을 경우 훈련을 종료하는데

이를 조기 종료라고 한다.

모형 설계에서 실제로 조기 종료가 어떻게 기능하는지는 다음과 같다. 모형을 개선한 후 drop-out rate에 의거해 삭제된 독립변수를 제외한 70%의 독립변수를 가지고 모형을 통해 예측을 한 후, 만일 개선한 모형의 설명력이 이전 모형에 비해 떨어진다면 stopping point로 1을 부여한다. 이렇게 해서 미리 설정해 둔 stopping point에 도달하면 프로그램 학습을 멈추고 해당 시점에서 최적의 모형을 결정한다. 만일 조기 종료를 5로 설정했다면, 모형을 개선하면서 이전 모형과 새로운 모형을 비교하여 이전 모형이 더 나은 경우 이전 모형을 새로운 모형으로 대체하지 않고 가지고 가되 stopping point 1을 적립하며, 새로운 모형에서 예측력이나 설명력이 떨어진 데 대해서 back-propagation을 통해 모형을 개선하고 새로운 모형으로 업데이트한다. 만일 다음 모형도 기존 모형보다 설명력이나 예측력이 좋지 않다면 모형은 개선하되 기존 모형은 추가적으로 stopping point를 적립하여 2가 된다. 이러한 식으로 미리 설정된 조기 종료 지점에 도달하면 해당 모형을 최적 모형으로 내놓는다. 하지만 중간에 새롭게 업데이트한 모형이 기존 모형에 비해서 더 나은 결과를 보여준 경우에는 최적 모형을 교체하며, 새로운 모형은 stopping point 0에서 시작해서 다시 미리 정해진 조기 종료에 도달할 때까지 새로운 모형과 비교된다.

조기 종료의 횟수가 높아지면 높아질수록 설명력은 높아지는 반면 학습 시간이 길어지고 과적합될 가능성도 올라간다. 반면 조기 종료를 위해 설정한 횟수가 짧아지면 모형의 일반성은 올라가나 적합도가 낮아진다. 대신 학습 시간이 크게 줄어드는 장점이 있다.

이러한 설정하에서 월단위로 정리된 데이터를 이용하여 모형은 학습을 한 후 노동시장 지표를 예측한다. 하지만 예측된 노동 지표의 값들은 실제값과 차이가 있을 수밖에 없다. 기계학습에서는 손실함수를 사용하여 예측값과 실제값의 차이를 계산하고 이 차이를 가지고 기계학습 모델의 매개변수들을 반복적으로 업데이트하여 예측값과 실제값이 근사해지도록 모형을 개선한다.

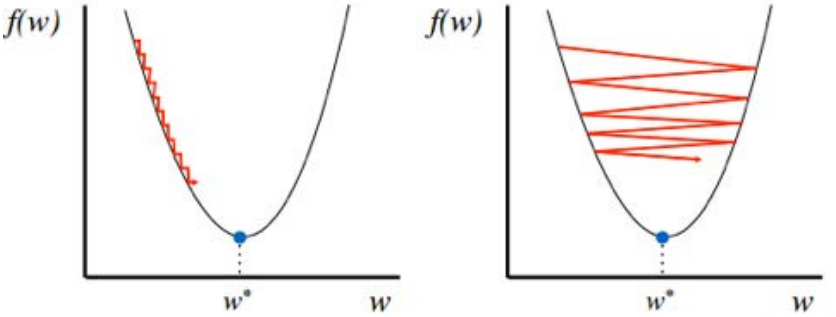
본 연구에서는 예측하고자 하는 변수들은 연속성이 있는 값들이므로,

손실함수로는 일반적으로 사용되는 평균 제곱 오차(mean squared error)를 사용하였다. 평균 제곱 오차를 활용하여 예측값과 실제값이 근사해지도록 모형을 계속해서 업데이트하는데, 이러한 업데이트 과정에도 여러 알고리즘이 쓰인다. 최근 심층학습 연구에서 가장 널리 쓰이는 알고리즘은 Adam(Adaptive Moment Estimation) optimizer와 SGD(Stochastic Gradient Descent)이다. Adam은 SGD에 비해서 성능이 크게 떨어지지는 않으면서도 훈련 및 학습에 소요되는 시간을 절약해 주는 특징이 있다. 연구에서 비정형 데이터를 사용하기 때문에 소요되는 긴 학습 시간을 고려하여 여기서는 Adam을 사용하였다.

손실 함수 및 조기 종료가 정의되었다면 마지막으로 프로그램에서 학습률(learning rate)을 정의해야 한다. 학습률이란 각 학습 사례들을 통해 모형을 얼마나 개선시킬 것인지 설정해 주는 것이다. 즉, 어떠한 모형을 학습한 후 인공신경망이 해당 모형의 개선점을 발견하였을 때, 얼마나 빨리 혹은 많이 이러한 개선점을 모형에 반영할 것인지를 결정하는 변수로, 기계학습에서 적정한 학습률을 설정하는 것은 매우 중요하다.

만일 학습률이 적정 수치보다 낮게 정의된 경우 학습에 소요되는 시간이 과도하게 길어지는데, 이것이 [그림 2-3]의 좌측에 묘사되어 있다. 한편 학습률이 높게 정의될 경우 모형은 최적화에 실패하는 경우가 발생할 수 있는데, [그림 2-3]의 우측 그림이 이러한 경우를 보여주고 있다. 학습률이 높기 때문에 학습 데이터나 테스트 데이터에 따라서 모형이 과도하게 많이 변화하며, 그로 인해 다수의 모형이 stopping point를 적립하지 못하고 새로운 모형에 의해 대체되어 모형의 개선만 이루어질 뿐 최적화된 값에 도달하지 못한다. [그림 2-3]의 우측에서는 지속적으로 모형이 개선되면서 최적점인 w^*에 도달해 가는 것처럼 보이지만, 학습률이 과도하게 높게 설정된 경우에는 아래에서 출발하여 위쪽 방향으로 움직이는 반대의 방향으로 모형 개선이 이루어져서 학습이 진행되면서 w^* 근처에서 시작했던 모형이 점차 w^*에서 멀어지는 방향으로 개선이 이루어질 수도 있으며, 또한 w^*를 향하지 않고 개선 방향이 양측을 반복하여 시소처럼 이동하기만 하면서 전혀 개선이 이루어지지 않는 경우도 있다.

〔그림 2-3〕 학습률에 따른 최적화 양상



자료 : 저자 작성.

모형의 설계에 대해서 요약하면, 본 연구에서는 drop-out rate은 표준적으로 사용되는 0.3으로 설정하여 새롭게 모형을 개선한 후 주어진 설명변수들 중 30%를 임의로 삭제하고 나머지 70%만을 가지고 모형을 검증토록 하였다. 또한 조기 종료는 10회로 설정하였으며, Adam의 초기 학습률을 0.001로 정의하였다. 하지만 이 학습률은 실제 학습을 수행하면서 조정되어야 한다.

이러한 기계학습을 통해 시스템은 노동시장의 주요 변수들의 익월 값들을 예측한다. 예를 들어, 2020년 7월까지의 각종 정형 및 비정형 데이터를 활용하여 2020년 8월의 노동시장 지표들을 예측하며, 8월까지의 각종 데이터를 활용해서 9월의 값들을 예측하도록 하는 것이다. 이를 통해 정형 데이터에서 이용 가능한 최신의 5개월에 대해서 시스템이 예측토록 하였는데, 이때 우리가 실제로 알고 있는 노동시장 지표 전망치는 테스트 데이터로 주어져서 모형이 얼마나 잘 예측하고 있으며 성능은 어떠한지를 시스템 스스로가 검증하는 데 사용하도록 설정했다.

3. 비정형 데이터의 처리

기계학습에서 비정형 데이터를 처리하는 방법은 크게 네 단계로 구성된다. 앞서 비정형 데이터를 어떻게 형태소로 구분하였으며 전처리가 어떠한 방식으로 진행되는지 설명하였다. 이렇게 구분된 형태소로 구성된

비정형 데이터, 즉 신문 기사 제목은 기계가 처리할 수 있는 수치화된 데이터로 작성된 자료가 아니므로 각각의 형태소마다 고유한 번호를 부여하는 과정을 먼저 거치게 된다. 이를 형태소 인덱싱(token indexing)이라고 부른다. 앞서 언급한 예시인 ['경제', '안', '좋다는데', '이상하네', '2100', '뚫은', '코스피', '3', '가지', '이유']의 10개의 형태소로 분할된 신문 기사 제목의 경우, 각각의 10가지가 모두 고유한 번호를 부여받는다. 다만, 다른 신문 기사 제목에 '경제'가 등장하는 경우에는 '경제'에 해당하는 형태소가 모두 같은 고유 번호를 공유한다.

이렇게 형태소에 고유의 인덱스를 부여하면, 이러한 인덱스를 가지고 신문 기사 제목을 숫자들의 나열로 변환할 수 있다. 이렇게 수치 혹은 숫자로 변환함으로써 비정형 데이터는 프로그램이 분석 가능한 정형 데이터로 바뀌게 된다. 이러한 숫자 변환 과정 이후 신문 기사 제목은 다음과 같은 형태의 4차원의 나열(4-dimension array)로 표시할 수 있다.

[[[[형태소] 기사별 타이틀]일별 전체 타이틀]월별 전체 타이틀]

여기서 형태소를 바탕으로 기사의 타이틀이 구성되며, 숫자로 구성된 기사의 타이틀을 바탕으로 해당 일의 전체 신문 기사 제목을 모아 일련의 숫자로 구성된 번호로 압축하게 된다. 앞서 언급하였듯이 우리가 예측하고자 하는 변수는 월별 자료이므로, 이렇게 구성된 일별 자료들을 모두 모아서 시스템은 이를 다시 월별 타이틀로 변환한다. 즉, 개별 신문 기사 제목을 모두 모아서 월별로 하나의 번호를 부여하며, 여기에는 개개의 신문 기사 제목의 성향과 논조, 주제 등이 모두 일련의 번호로 변경되어 담겨 있다.

그런데 이렇게 월별 타이틀로 변환된 형태소의 길이는 모두 각각 다르다. 그 길이가 다른 이유는 세 가지 이유 때문인데, 하나는 기사마다 제목의 길이가 달라서 기사 제목에 따른 형태소 인덱싱의 길이가 다르다. 또한 각 일자별로 신문 기사의 숫자가 다르기 때문에 이로 인해서 일별로 타이틀화된 신문 기사들에 대한 정보가 다르게 된다. 마지막으로 각 월별로 일자의 숫자가 다르기 때문에 월별 전체 타이틀의 길이가 다르다. 예를 들어, 2월은 28개 혹은 29개의 일별 정보만을 담고 있는 반면 4월이나

6월은 30개, 10월이나 12월은 31개의 일자를 가지고 있어, 설령 일별로 타이틀 길이가 같다 하더라도 월별 타이틀의 길이는 달라지게 된다. 그러나 심층학습 틀에서 타이틀을 처리하기 위해서는 모든 타이틀의 길이를 동일하게 만들어줘야 한다. 이 과정을 패딩(padding)이라고 한다.

패딩은 각 타이틀 단위별로 짧은 타이틀을 가장 긴 타이틀에 맞추어 길이를 늘린다. 이때 인위적으로 늘어나는 타이틀 길이에는 모두 0을 부여하여 이들이 패딩에서 인위적으로 늘려진 값이라는 것을 표시한다. 따라서 패딩 작업이 완료된 후 4차원 나열의 모양은 다음과 같다.

[maximum length among title tensors * maximum number of daily items * 31 * 238]

여기서 'maximum length among title tensors'는 가장 많은 형태소를 가진 신문 기사의 형태소 숫자이며, 'maximum number of daily items'는 가장 많은 신문 기사 제목이 생성되었던 날의 신문 기사 제목 숫자이다. 31은 12개월 중 가장 일수가 많은 달의 일수를 표시한 것이며, 238은 2001년 01월부터 2020년 10월까지의 개월 수를 지칭한다.

자연어 처리에서 기본적인 분석 단위는 형태소이다. 형태소는 식별 가능한 단어인 경우도 있고 경우에 따라서는 단어의 일부분인 경우도 있다. 같은 형태소라 하더라도 문맥에 따라서 그 의미가 상이할 수 있기 때문에 각 형태소의 의미를 하나의 숫자인 인덱스로 표현할 경우 그 형태소가 가지는 의미를 충분히 표현해 내기 힘들다. 따라서 각 형태소의 문맥상의 의미를 표현하기 위해 각 단어를 n차원의 벡터로 표현하게 되는데 이 과정을 형태소의 벡터화(vector representation)라고 하며 벡터화를 끝낸 형태소 단위의 결과값 벡터를 워드 임베딩(word embeddings)이라고 한다. 워드임베딩을 도출하는 방법으로 BERT, ELMo, Word2Vec, GloVE 등 다양한 모델들이 존재하는데, 어떤 모델에서건 각각의 형태소는 고유한 벡터값을 가진다. 이러한 워드 임베딩은 모델에 따라 50차원에서 1024차원에 이르는 값을 가진다.

이렇게 해서 패딩이 완료된, 4차원 나열로 표시된 비정형 데이터는 심층학습의 틀에서 처리하기 위해서 데이터 처리의 기본 단위인 텐서(tensor)로 변환한다.

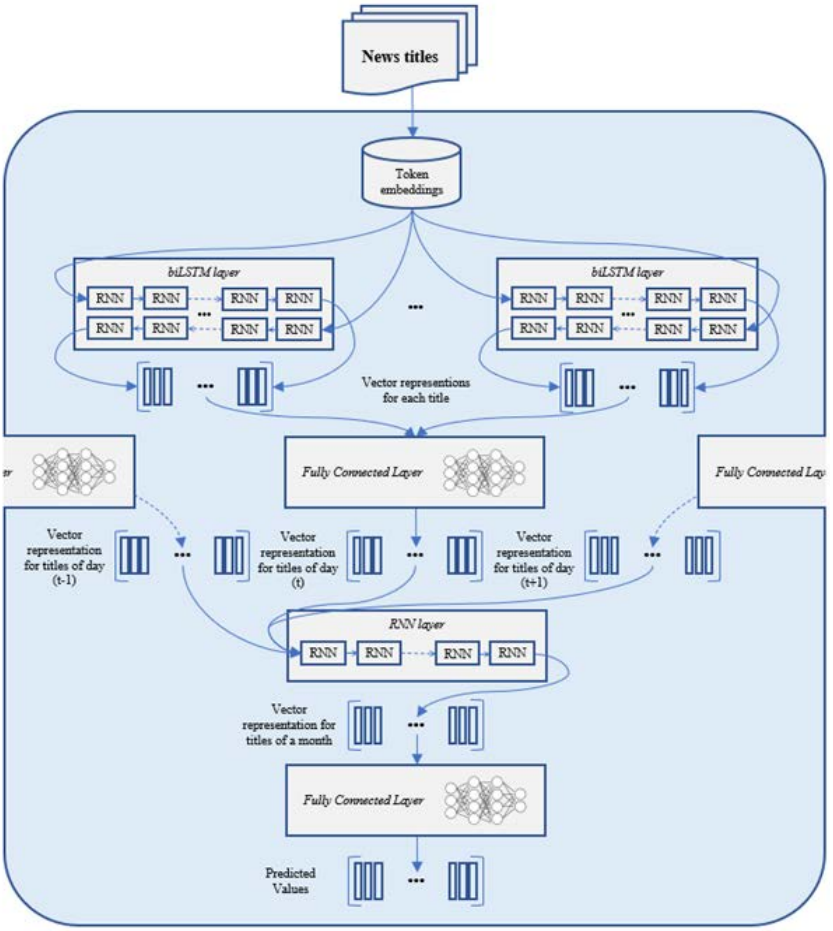
4차원 나열 값으로 변환된 텐서의 각 요소들에 100차원의 고유 벡터값이 할당되면 텐서는 이제 5차원의 텐서가 되며 [maximum length among title tensors * maximum number of daily items * 31 * 238 * 100]의 모양으로 바뀐다.

벡터화된 형태소를 처리하는 딥러닝 유닛에는 여러 종류가 있는데, 크게는 각 요소의 나열 순서가 의미를 가지느냐 그렇지 않느냐에 따라 적합한 유닛이 달라진다. 다시 말해서 3, 4, 5, 6이라는 벡터값이 나열되는 순서에 따라 의미가 다르냐, 혹은 순서와 무관하게 같은 의미를 가지느냐에 따라 사용해야 하는 유닛이 달라지는데, '성장률 하락 실업률 증가'는 '성장률 증가 실업률 하락'과 정반대의 의미를 가진다. 일반적으로 각 요소의 순서 역시 의미를 가지며 그에 맞추어 해석하는 자연어의 특성상 입력값을 순차적으로 처리하는 순환 신경망(Recurrent Neural Networks : RNN)이 자연어의 처리에 적합하다고 여겨진다. 순환 신경망에는 LSTM 또는 GRU 등 메모리 기능을 가지고 있어 현재 처리되고 있는 입력값과 멀리 떨어진 입력값의 연관성을 발견해 낼 수 있는 순환 신경망과 단순 순환 신경망(Vanilla RNN)이 있는데, 본 연구에서는 메모리 기능이 있는 LSTM을 분석 유닛으로 사용하였다.

형태소를 통해서 LSTM으로 읽어낸 결과물은 건별 신문 기사 제목을 나타내는 벡터값이 된다. 건별 기사 제목을 표현하는 벡터는 일별로 모여져서 완전 접속망(fully connect networks)을 통과하게 되고 일별 기사 제목 전체를 대변하는 벡터로 변환된다. 일별 신문 기사 제목은 다시 시간의 순서에 따라 LSTM 유닛을 통과하게 되고, 이 LSTM을 통과한 결과물은 월별 기사 제목을 대표하는 벡터가 된다. 최종 완전 접속망에서는 이 월별 신문 기사 벡터를 이용하여 월별 노동지표를 예측하게 된다. 이러한 처리과정은 [그림 2-4]에 시각화되어 있다.

앞서 정형 데이터에서 일부 금리 및 주가 관련 통계들은 일별로 제공된다고 하였다. 이러한 일별 정형 데이터 역시 일별로 제공되는 신문 기사 제목과 마찬가지로 순환 신경망을 거쳐서 월별 벡터로 변환한 후 예측 모형에 삽입한다.

〔그림 2-4〕 심층학습 시스템 개요도



자료 : 저자 작성.

제4절 소 결

본 장에서는 연구에 사용할 기계학습 및 심층학습의 원리와 다양한 방법론을 살펴보았다. 심층학습이 범용화되어 자연어 처리가 가능해진 것은 이후 많은 학문에서 연구 분야를 확장하는 데 큰 기여를 할 것이다. 기

존에는 컴퓨터가 바로 인식하지 못했기 때문에 사용이 힘들었거나 혹은 많은 시간과 노력을 들여야만 분석할 수 있었던 영역이 충분한 컴퓨팅 자원만 있으면 처리 가능하게 되었기 때문이다. 또한 그간 수치화된 통계만을 사용하다보니 분석 가능한 자료의 범위가 제한되었으나, 자연어 처리 기법은 이러한 제약을 획기적으로 완화시켜 주었다.

본 연구에서는 비정형 데이터를 이용하여 노동시장의 주요 변수들인 경제활동참가율, 고용률, 실업률을 예측해 보고자 한다. 예측을 위해서 사용하는 자료는 크게 세 가지로 나눌 수 있는데, 첫째는 노동시장 관련 변수들로, 이들을 이용한 분석을 통해 노동시장에 강한 내생성이 존재하는지 여부를 알아볼 것이다. 그리고 두 번째 모형은 노동시장 통계 이외의 모든 접근 가능한 일별 및 월별 자료들을 사용할 것인데, 여기에는 비단 실물 경제의 각 분야에 대한 통계치들뿐만 아니라 각종 사회경제적 지표들, 경제 주체들의 기대를 반영한 전망 지수들과 각종 인구 통계 등을 망라한다.

마지막 모형에서는 비정형 데이터를 사용하였는데, 분석에서 사용한 비정형 데이터는 지난 20여 년간의 중앙 일간지 및 주간지 신문 기사 제목 전체이다. 자연어는 분석을 위해서 여러 전처리 과정을 요하는데 이러한 전처리 과정에서 최대한 연구자의 임의성과 주관을 배제하기 위해 최소한의 개입만 수행했다.

이렇게 전처리를 거친 비정형 자료는 형태소 단위로 수치 변환한 후, 다시 최종적으로 5차원 배열의 텐서로 변환된다. 이때 단어의 나열 순서나 기사의 순서가 경기 변동의 흐름을 설명하는 의미를 가지기 때문에 순차적인 배열로부터 의미를 읽어내는 순환 신경망 방법을 통해 비정형 자료와 일별 통계 자료는 월 단위 수치 벡터로 전환된다. 이렇게 일별 자료들을 월 단위로 변환하면 모든 자료들이 예측하고자 하는 노동시장 지표와 마찬가지로 월 단위 자료가 되며, 이를 가지고 모형은 경제활동참가율, 실업률, 고용률을 예측한다. 이때 예측값과 실제값은 차이가 있는데, 모형에 대한 평가는 두 값의 차이를 이용하여 평균 제곱 오차를 계산한 후, 이러한 평균 제곱 오차를 최소화하도록 모형을 개선하게 설계하였다.

기계학습을 통해 설계된 모형의 성과를 측정하는 지표는 크게 일반성

과 적합도가 있는데, 일반성은 모형이 개개의 변수에 크게 영향을 받지 않고 얼마나 추세적으로 실제값을 정확하게 예측하는가 하는 지표이며, 적합도는 주어진 모든 설명변수를 통해 실제값을 최대한 정확하게 설명하도록 모형을 설계하는 지표이다. 적합도와 일반성 간에는 상충 관계가 존재하는데, 이 두 지표를 잘 충족시키는 최적의 모형을 찾기 위해서 모든 기계학습 모형에 대해서 미리 설정된 drop-out rate에 기초하여 독립변수 중 일부 값을 삭제하고 성과를 측정한다. 이를 통해, 개선되는 모형들이 기존 모형보다 성과가 떨어져서 미리 설정한 stopping point가 충족되면 기계학습은 최적화된 모형을 결정한 후, 해당 모형에 기초하여 노동시장 변수들을 예측한다.

제 3 장
기계학습 모형의 예측 결과

제1절 예측 모형 I

1. 모형 설정 및 개요

첫 번째로 학습시켜 예측값을 얻어볼 모형은 노동시장 변수만을 사용한 모형이다. 모든 시장은 – 상품 시장과 서비스 시장을 포함하여 – 일정 부분 내생성이 존재한다. 만일 노동시장에 강한 내생성이 존재하며, 노동시장의 변동성 대부분이 계절적 요인 및 노동시장 내에서 관찰되는 장기 및 단기의 추이에 의해 결정되는 성격이 강하다면, 노동시장 지표에 대한 예측은 과거 노동시장 변수들만 사용하는 것으로도 충분할 것이다.

또한 한국의 노동시장이 경직적인 구조를 가지고 있다면, 외부 요인의 변화에 의해 크게 영향을 받지 않을 수 있다. 즉, 해고가 어려운 경우 호경기에도 채용 인원을 쉽게 늘리려고 하지 않을 것이며, 불경기에는 기업의 경영 성과가 위축됨에 따라 인원을 조정하여 인건비를 줄이는 선택을 하지 못하고 계속해서 고용량을 유지해야 할 수 있다. 물론 비정규직의 비율이 높은 우리나라 노동시장의 특성을 고려하면 이들까지 고려한 노동시장 통계에서 노동시장의 경직성이 얼마나 유효하게 작용할지는 알 수 없다.

앞서 정형 데이터에서 소개했듯이 노동시장에 관한 통계로는 통계청에서 제공하는 경제활동인구조사의 성별, 연령별, 교육정도별 경제활동인구 자료를 활용하였다. 이를 통해 성별, 연령대별, 교육수준별로 경제활동참가율이나 실업률, 고용률이 지난 20여 년간 어떻게 변화해 왔는지와 함께 계절적인 요인 및 최근의 추세 등을 고려하여 익월의 이들 변수 값들을 예측할 것이다.

이 데이터를 통해 노동시장에 참가하는 사람들이나 취업자 및 실업자들의 성별 분포, 연령별 변화 추이, 그리고 지난 20여 년간의 평균적인 학력수준 및 개별 학력수준별 노동시장에서의 활동 지표들이 모두 포함되어 있어 노동시장과 관련된 여러 변수들의 단기, 중기 및 장기 추세들과 경제활동인구 및 취업자들의 요인별 분포 변화까지 반영할 수 있다. 이를 통해 사회 전체의 경제활동 가능인구의 특성과 추이를 추려내고, 계절적 요인에 의한 변화를 포착한 후 이를 반영하여 노동시장 지표들을 예측한다.

이 모형의 장점은 노동시장에 직접적으로 연관된 변수들만을 포함하고 있다는 것이다. 만일 노동시장이 경직적이거나 다른 재화 및 서비스 시장과 많은 상관관계를 가지지 않는다면 다른 변수들은 노이즈로 작용하여 예측을 방해할 수 있다. 본 모형에서는 노동시장 이외의 변수들을 고려함으로써 다른 노이즈가 대부분 제거되어 예측의 정확성이 제고될 수 있다. 하지만 노동시장이 금융시장이나 일부 상품 및 서비스 시장의 생산이나 재고 지수, 혹은 수출입 지표와 연관되어 있다면 모형에서 이들 변수를 활용할 수 없기 때문에 예측력이 떨어질 수 있다.

또한 시장 외적인 요인에 의해 경제 전반에 큰 변동성이 나타나거나 혹은 경제적 요인 이외의 충격, 예를 들면 전쟁이나 대규모 인구 이동, 전염병 등이 발생한 경우, 노동시장 변수들에 이러한 요인들의 영향이 반영되기에는 긴 시간이 소요되는 반면, 여타의 정형 데이터는 발생 직후 길지 않은 시간 동안에 해당 사건이 미치는 영향의 크기를 비교적 정확하게 반영하여 선제적으로 노동시장 변수의 변화를 예측케 할 수 있다. 신문기사로 구성된 비정형 데이터의 경우에는 이러한 경제 외적인 충격을 거의 실시간으로 반영하여 이를 예측 모형에 반영할 수 있다. 하지만 노동시장 지표만을 사용하는 모형에서는 이러한 충격이 노동시장에 반영된

이후에야 비로소 그 효과의 방향과 크기를 알고 뒤늦게 반응하게 된다는 약점이 있다.

예측은 예측 모형을 설계하도록 학습시킨 후, 2020년 6월까지의 자료를 가지고 7월을 예측하고, 7월까지의 자료를 활용하여 8월을, 8월까지의 자료로 9월 노동시장 지표를, 9월까지의 통계를 이용하여 10월을 예측하고, 10월까지의 지표를 이용해서 11월 지표까지, 5개월을 예측한 후 실제값과의 차이를 비교토록 하였다.

2. 예측 결과

노동시장의 변수만을 활용하여 익월의 노동시장 변수를 예측케 한 모형의 예측 결과는 <표 3-1>에 제시되어 있다. 세 가지 변수 중에서는 경제활동참가율에 대한 예측력이 가장 좋으며, 고용률에 대한 예측력이 그 다음이고, 실업률에 대한 예측력이 가장 떨어지는 것으로 나타났다.

실업률의 경우 특히 2020년 7월에서 2020년 8월로 갈 때 0.9%p가 떨어져서 크게 하락했다가 2020년 9월에 3.6%로 반등하였는데, 예측 모형에서는 이를 정확하게 예측하지 못하고 2020년 8월에는 도리어 7월보다 실업률이 높아지는 것으로 예측했다. 그러다가 9월 지표부터 실업률이 급격

〈표 3-1〉 예측 모형 I의 예측 결과

	실제값			예측치		
	경제활동 참가율	실업률	고용률	경제활동 참가율	실업률	고용률
2020년 7월	63.1	4.0	60.5	63.3	4.0	60.5
2020년 8월	62.4	3.1	60.4	62.8	4.2	60.1
2020년 9월	62.5	3.6	60.3	62.2	3.3	60.1
2020년 10월	62.7	3.7	60.4	62.7	3.4	60.6
2020년 11월	62.8	3.4	60.7	62.3	3.8	59.9
평균 제곱 오차				0.09976	0.29212	0.14773

자료 : 저자가 설계한 모형의 결과.

히 낮아지는 값을 예측하였으나, 실제 9월에는 실업률이 크게 상승하였으며 이후에는 비슷한 값을 보이는데, 예측치는 직전 수개월간의 큰 변동폭이 11월까지 계속된다고 전망한 것으로 보인다.

평균 제곱 오차로 측정하였을 때 경제활동참가율이 가장 예측의 정확성이 높았으며 고용률이 그다음인데, 고용률의 경우 예측 기간 동안 상당히 안정적으로 60.4% 근처를 유지한 반면 경제활동참가율은 62.4%~63.1% 범위에서 상대적으로 움직임이 컸다. 가장 움직임이 큰 변수는 실업률이었는데, 노동시장 변수들이 변화폭이 크지 않은 편이기 때문에 오히려 해당 기간에 급격한 변동성을 띠었던 실업률의 경우 예측력이 크게 떨어졌으며, 고용률의 경우 주어진 기간 동안 변화가 거의 없는 움직임을 보였기 때문에 경제활동참가율과 비교하여 예측력이 저하된 것으로 보인다. 즉, 모형은 노동시장 변수들이 단기간에는 직전 수개월 값들의 평균 근처에서 0.5%p의 편차로 움직일 것이라고 예측한 것으로 보이지만, 실업률의 경우에는 변화가 과도하게 커서, 고용률은 변화가 거의 없어서 경제활동참가율에 비해 예측의 정확도가 낮았던 것으로 보인다.

기계학습 및 심층학습을 활용하여 예측할 때 중요한 지표 중 하나가 예측하고자 하는 지표의 변화 방향을 정확하게 맞추었는가이다. 사실 수

〈표 3-2〉 예측 모형 I의 변화 방향 예측 결과

	실제값			예측치		
	경제활동 참가율	실업률	고용률	경제활동 참가율	실업률	고용률
2020년 7월	−	−	+	+	−	+
2020년 8월	−	−	−	−	+	−
2020년 9월	+	+	−	−	+	−
2020년 10월	+	+	+	+	−	+
2020년 11월	+	−	+	−	+	−
정확도				2/5	2/5	4/5

자료 : 저자가 설계한 모형의 결과.

치를 예측하는 것은 100% 일치하지 않을 수 있기 때문에 많은 경우 모형의 정확성을 재는 한 측도로 모형의 예측값이 가리키는 변화 방향이 실제 값의 변화 방향과 같은지를 살펴본다. 노동시장 지표들의 경우 지표가 감소하였는지 증가하였는지 그 자체도 큰 의미를 가지기 때문에 변수의 증감 여부를 맞추는 것은 모형의 성능 측면에서 의의가 있다.

앞서 평균 제곱 오차를 통해 수치가 얼마나 근접했는지를 살펴본 결과에서는 경제활동참가율의 정확성이 가장 높고, 그다음이 고용률, 그리고 실업률이 가장 낮은 예측력을 보이는 것으로 분석되었다. 하지만 변화 방향만 놓고 보면 모형 I은 고용률이 움직이는 방향을 예측하는 데 있어 상당히 높은 정확성을 보였다. 고용률의 경우 실제로는 2020년 11월에 증가하였지만 모형에서는 감소한 것으로 예측한 것을 제외하면 모든 경우에 고용률의 증감 여부 자체는 정확하게 예측하였다. 반면 경제활동참가율과 실업률의 경우에는 증감 여부를 상대적으로 정확하게 예측하지 못하여 수치상으로는 근접한다 하여도 노동시장 전반의 변화 방향은 정확하게 파악하지 못한 것으로 나타났다.

이러한 변화 방향이 중요한 이유는 앞서 언급했듯이 경제활동참가율, 실업률, 고용률 지표가 월별로 크게 변화하는 값들이 아니라는 특성 때문이다. 만일 예측의 정확성을 높이기 위해서 모형이 직전 몇 개월의 평균값만을 계속해서 제시하거나 중간값만 제시한다면 평균 제곱 오차로 측정한 예측력은 높을 수 있다. 하지만 이는 노동시장 지표 변화가 크지 않기 때문에 우연히 정확성이 높게 측정된 것이며, 실제 노동시장이 어떻게 변화하는지는 전혀 예측하지 못한다고 해석할 수 있기 때문이다.

상술한 두 가지 예측력 지표를 보았을 때, 모형 I이 경제활동참가율, 실업률, 고용률 중에서 어떠한 지표를 예측하는 데 강점이 있다고 말하기는 힘든 결과를 얻었다. 평균 제곱 오차로 계산한 수치상으로는 경제활동참가율의 예측력이 가장 좋았으나 변화 방향만을 놓고 보면 고용률에 대한 예측력이 월등히 좋았기 때문이다. 따라서 실업률의 경우 노동시장 변수들 중 내생성이 가장 약하며, 노동시장 변수만으로 정확하게 예측하는 것이 쉽지 않은 변수라고 추정할 수 있다.

제2절 예측 모형 II

1. 모형 설정 및 개요

두 번째로 학습시키는 모형에서는 모든 정형 데이터를 포함하여 예측하도록 설계하였다. 노동시장이 다른 재화 시장이나 서비스 시장과 깊은 상관관계를 가지고 있거나 금융시장의 추이에 따라서 기업의 경제활동과 노동 수요가 크게 영향을 받는다면, 모든 정형 데이터를 활용한 모형이 가장 정확한 예측값을 줄 수 있다. 특히나 해외 부문이 차지하는 비중이 경제의 반 이상인 한국 경제의 특성상 해외 부문에서의 움직임을 대변하거나 혹은 해외 시장과의 경제활동 정도를 측정하는 각종 수출입 자료가 포함된 정형 데이터는 예측력이 높을 수 있다.

만일 노동시장이 내생성이 아주 크지 않으며 경제의 다른 부분들과 유기적으로 연결된 관계를 가지고 있다면, 모형 II는 해당 부분들까지 반영하여 예측력이 높아질 수 있다. 예를 들어, 각종 제조업이나 서비스의 생산 지수는 해당 시기에 경제활동이 얼마나 활발한지를 측정한다. 모형이 정확하게 어떠한 모델을 세워서 예측하는지는 기계학습의 특성상 알 수 없지만, 경제활동의 정도가 시차를 두고 노동시장에 영향을 준다 하더라도 모형은 일반적으로 시차 효과까지 고려하기 때문에 시차를 두고 나타나는 노동시장의 변화도 포착할 수 있다.

경제 주체들이 경제활동을 하거나 노동시장에 참여할지 여부를 결정할 때 주요하게 참고하는 변수가 금리와 관련된 변수인데, 본 모형에서는 각 월 단위로 일자별 금리의 변화 추이까지 텐서화하여 포착하므로 경제학에서 다루는 경제활동 결정 요인들을 대다수 포함하고 있다. 또한 경제주체들이 실물 시장의 움직임에 따라서 구직 활동이나 구인 활동의 정도를 조절하거나 그에 영향을 받는다면 모형 II는 이러한 움직임을 가장 정확하면서도 편의 없이 담을 수 있기 때문에 모형의 정확성이 올라갈 수 있다.

다만 모형 I과 비교하여 노동시장과 무관한 많은 변수들 역시 포함되어 있을 가능성이 높기 때문에 이들을 얼마나 노이즈로 처리할 수 있는지에 따라 모형의 예측력은 크게 달라질 수 있다. 만일 노이즈를 효과적으로 잡아내지 못하거나 혹은 활용 가능한 변수가 증가하면서 과적합한 모형을 만들어내면 오히려 익월 노동시장 변수에 대한 예측력은 떨어질 수 있다. 특히 모형에서 사용하는 변수들의 총 개수가 12,027개에 이르러 이들 중 실제 노동시장과 깊이 연관되어 있거나 설명력이 높은 변수가 얼마나 되느냐, 그리고 노이즈인 변수와 그렇지 않은 변수들을 얼마나 잘 식별하여 모형을 설계하였느냐에 따라 모형의 예측력이 영향을 받을 수 있다.

일반적으로 기계학습과 심층학습에서는 예측에 사용되는 자료의 크기가 크면 클수록, 다시 말해서 데이터가 많아지면 많아질수록 예측력이 올라간다고 본다. 하지만 이는 예측하고자 하는 변수와 연관 있는 관측치들이 많을 경우에 그러하며, 노이즈가 많은 것이 예측력을 반드시 높이지는 않는다. 따라서 모형 I과 모형 II의 결과를 비교하여 본다면 노동시장이 얼마나 내생성이 큰지, 그리고 노동시장이 얼마나 경직적인지를 간접적으로 살펴볼 수 있을 것으로 생각한다.

2. 예측 결과

모든 정형 데이터를 활용하여 익월의 노동시장 변수를 예측한 결과는 <표 3-3>에 제시되어 있다. 평균 제곱 오차로 측정하였을 때 앞서 <표 3-1>에서와 마찬가지로 세 가지 변수 중에서는 경제활동참가율에 대한 예측력이 가장 좋고, 고용률을 그다음으로 정확하게 예측하며, 실업률에 대한 예측력이 가장 낮은 것으로 나타났다. 앞서와 마찬가지로 실업률이 2020년 8월을 경계로 직전, 직후 월에 전혀 다른 방향으로 크게 변하였는데, 이에 대한 예측에 있어서 모형의 예측 정확성이 크게 떨어지는 것으로 나타났다.

경제활동참가율에 있어서는 모형은 실제보다 더 변동폭이 크게 움직이는 것으로 예측하여 예측력이 낮아지는 것으로 나타났다. 반면 실업률의 경우 초기 3개월을 제외하면 마지막 2개월의 경우 비교적 근사하게 예

〈표 3-3〉 예측 모형 II의 예측 결과

	실제값			예측치		
	경제활동 참가율	실업률	고용률	경제활동 참가율	실업률	고용률
2020년 7월	63.1	4.0	60.5	63.1	4.4	60.3
2020년 8월	62.4	3.1	60.4	62.7	4.1	60.0
2020년 9월	62.5	3.6	60.3	62.0	3.3	60.1
2020년 10월	62.7	3.7	60.4	62.7	3.4	60.7
2020년 11월	62.8	3.4	60.7	62.2	3.5	59.9
평균 제곱 오차				0.11976	0.30348	0.18229

자료 : 저자가 설계한 모형의 결과.

측하여 실업률의 변동폭이 이례적으로 크지 않다면 예측력이 나아질 수 있음을 시사한다. 마지막으로 고용률의 경우 비교적 근사한 예측을 하다가 2020년 11월 값에 대해서 전월과 비교하여 크게 떨어질 것으로 예측하였는데, 실제로는 소폭 상승하여 이로 인해서 평균 제곱 오차가 크게 상승했다.

지표의 변화 방향에 대한 정확성은 <표 3-4>에 제시되어 있다. 앞선 모형 I에서는 고용률의 변화 방향은 정확하게 맞추되 나머지 두 변수에 대해서는 크게 틀린 예측을 하였으나, 모형 II는 모든 경제활동참가율과 고용률에 대해서 비교적 평균적인 예측력을 보인 반면, 역시 실업률에 대해서만 매우 낮은 예측력을 보였다. 여러 실물 경제 변수들을 포함하다보니 모형 II는 노동시장에 영향을 미치는 여러 변수들의 효과가 모두 고려되어 상대적으로 안정적인 움직임을 보이는 경제활동참가율과 고용률에 있어서는 모두 비교적 준수한 예측력을 보인 것이 특징이다.

하지만 변화 방향에 대한 전반적인 정확성은 앞선 모형 I과 비교하여 전체적으로 크게 다르지 않다. 오히려 평균 제곱 오차의 크기가 모든 변수에 대해서 더 크기 때문에 모든 정형 데이터를 포함하여 예측하는 것은 노이즈로 작용하는 변수가 많아 예측력을 떨어뜨릴 뿐이며, 오히려 내생

〈표 3-4〉 예측 모형 II의 변화 방향 예측 결과

	실제값			예측치		
	경제활동 참가율	실업률	고용률	경제활동 참가율	실업률	고용률
2020년 7월	−	−	+	−	+	−
2020년 8월	−	−	−	−	+	−
2020년 9월	+	+	−	−	+	−
2020년 10월	+	+	+	+	−	+
2020년 11월	+	−	+	−	−	−
정확도				3/5	2/5	3/5

자료 : 저자가 설계한 모형의 결과.

성이 강하거나 노동시장 변수와 직접적으로 연결된 변수들만 예측에 활용하는 것이 모형의 정확성을 높일 수 있음을 시사하고 있다. 하지만 전반적인 변수의 움직임을 측정하는 데에서는 두 모형이 큰 차이가 없는 것을 통해, 경제활동참가율은 다른 실물 변수들과도 깊게 연결되어 있으며, 고용률의 경우에도 실물 변수까지 고려하는 것이 예측력을 크게 저하시키지 않는다는 결론을 얻었다.

제3절 예측 모형 Ⅲ

1. 모형 설정 및 개요

마지막 예측 모형은 모든 정형 데이터에 신문 기사 제목까지 포함하여 예측하는 모형이다. 심리나 여론이 실물 경제 변수에 미치는 영향이 얼마나 큰지에 대해서는 경제학계에서 여전히 갑론을박이 있으나, 적어도 특정한 상황에서는 영향을 미친다는 것이 대체적인 의견이다. 예를 들어, 대공황 시기의 뱅크런(bank run; 급작스런 대규모 은행 예금 인출 사태)은 사람들의 심리와 사회 분위기가 다수의 경제활동을 한 방향으로 유도

하여 금융 시장, 최종적으로 실물 경제 전체에까지 영향을 미친 바 있다.[3)]

노동 수요나 공급이 사회 분위기나 여론에 의해 영향을 받는다면 여론의 동향을 살필 수 있는 신문 기사를 포함하는 것이 예측력을 높이는 데 도움이 될 것이다. 또한 역으로 신문 기사가 특정한 방향으로 일어나는 경제활동의 동향에서 나타나는 특징을 보다 잘 잡아낼 수도 있다. 즉, 거시적인 지표에서는 포착되지 않는 특정한 집단이나 계층, 혹은 산업 부문에서 나타나는 경제적 변화를 여론에서 크게 다룬다면, 그러한 부문의 영향이 정형 데이터에서는 포착되지 않는다 하더라도 비정형 데이터에서는 잡힐 수 있다. 더하여 경제 주체들의 행동이 실제 실물 경제 통계에 반영되는 데 시차가 소요될 수 있으나, 여론은 그러한 변화를 즉각적으로 반영할 수 있는 장점도 있다.

빅데이터 분석의 특징은 보다 많은 데이터가 보다 좋은 예측력이나 설명력을 낸다는 것인데, 앞선 모형 Ⅰ이나 Ⅱ와 비교하여 모형 Ⅲ은 비정형 데이터까지 포함하므로 데이터의 크기만을 놓고 보면 비정형 데이터의 크기와 양이 가장 많다. 따라서 활용하는 자료 역시 모형 Ⅲ이 가장 많으므로 일반적인 심층학습에서는 가장 예측력이 높을 수 있다. 하지만 이는 양날의 검과 같아서 앞서 모형 Ⅱ에 대한 설명에서 언급했듯이 노이즈처럼 작용하는 것을 배제할 수 없다.

더하여 노동시장과 관련되어 있지 않은 신문 기사만을 특정할 수 없기 때문에 노동시장과 무관한 다수의 기사들이 포함되어 있다는 점은 학습 데이터에 많은 노이즈를 삽입하여 예측력을 크게 낮추는 역할을 할 수 있다. 하지만 이를 위해서 경제 관련 기사나 사회 관련 기사만을 추리는 것도 문제인 것이, 경제나 사회 기사들 중에서도 노동시장과 무관한 것들이 많을 뿐만 아니라, 각종 스포츠 행사나 정치적 의사결정에 따른 노동시장에 적용되는 법률이나 정책 방향의 변화, 정부와 입법 기관의 노동시장에 대한 태도 등은 경제나 사회면에 포함되지 않는데 이들을 배제한다면 노동시장에 영향을 미칠 수 있는 여론이나 언론의 행위를 모두 포착하지 못

3) Bernanke, B. S.(1983), “Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depressio.” *American Econ Review* 73 (3), pp.257~276.

한다는 문제로 인해 마찬가지로 예측력이 떨어진다.

모형 Ⅲ에서 비정형 데이터를 포함했을 때 예측력이 낮아질 수 있는 또 다른 요인으로는 쏠림 현상이나 특정 사회 현상만이 부각될 우려가 있다는 점이다. 신문 기사들은 대체적으로 당시 사회적으로 크게 논란이 되거나 논쟁이 되는 영역을 많이 다루는데, 이 경우 해당 주제의 영향이 과도하게 반영될 우려가 있다는 점이다. 반대로 노동시장에 영향을 미치는 정책 변화라 하더라도 사회적으로 문제나 논쟁의 대상이 되지 않는다면, 이러한 것들이 여론에서는 충분한 양만큼 다뤄지지 않아 과소추계될 여지가 있다.

다만, 앞서 확인했듯이 모형 Ⅱ에서의 예측력이 모형 Ⅰ과 비교하여 크게 높아지지 않는 것으로 나타나, 1,000여 개가 넘는 실물 경제 변수들에서 있을 수 있는 편의나 왜곡을 비정형 데이터가 바로잡아 예측력이 개선될 수도 있다.

2. 예측 결과

<표 3-5>는 모든 정형 데이터와 비정형 데이터를 활용하여 익월의 변수를 예측한 결과를 보여주고 있다. 평균 제곱 오차로 살펴보면 실업률과 고용률에 대해서는 모형 Ⅲ의 예측력이 가장 낮은 것으로 나타났으며, 경제활동참가율의 경우에는 모형 Ⅰ보다는 예측력이 떨어지지만 모형 Ⅱ보다는 나은 것으로 나타났다.

개별 노동시장 지표들에 대한 평균 제곱 오차에서의 특징으로는 역시 다른 모형과 마찬가지로 실업률에 대한 예측력이 크게 떨어진다는 점이다. 실업률에서는 8월에 크게 하락한 추세가 9월에 반영되었으며, 9월에 반등한 것은 11월에서야 나타나는 등 일정한 시차를 두고 실제 실업률의 움직임을 추종하여 예측하는 모습을 보이고 있다.

한편 고용률의 경우, 실제 변수가 큰 변동을 보이지 않지만, 모형에서는 고용률이 크게 하락하는 모습이 두 차례 나타나는데, 이러한 하락이 평균 제곱 오차를 크게 늘리는 요인으로 작용한 것으로 보인다. 경제활동참가율은 마지막 11월에서 크게 하락한 것을 제외한다면 대체로 실제값

〈표 3-5〉 예측 모형 III의 예측 결과

	실제값			예측치		
	경제활동 참가율	실업률	고용률	경제활동참 가율	실업률	고용률
2020년 7월	63.1	4.0	60.5	63.2	4.2	60.7
2020년 8월	62.4	3.1	60.4	62.7	4.2	60.1
2020년 9월	62.5	3.6	60.3	62.1	3.2	60.1
2020년 10월	62.7	3.7	60.4	62.8	3.5	60.7
2020년 11월	62.8	3.4	60.7	62.3	3.8	59.8
평균 제곱 오차				0.1062511	0.311765	0.2155308

자료 : 저자가 설계한 모형의 결과.

과 큰 차이가 없게 예측하여 평균 제곱 오차가 셋 중 가장 낮은 값을 기록하였다.

예측한 지표들의 변화 방향이 정확한가에 대한 결과는 <표 3-6>에 제시되어 있다. 눈에 띄는 특징은 실업률의 경우 단순히 평균 제곱 오차만 큰 것이 아니라 예측치의 증감 여부와 실제값의 증감 여부가 크게 다르다는 것이다. 이는 실제값에서 실업률이 2020년 8월을 기점으로 앞뒤로 크게 변화한 것을 정확하게 예측하지 못한 점도 크지만, 전반적으로 실업률의 변화에 있어서만큼은 모형 III의 예측 정확성이 크게 떨어진다는 것을 다시 한 번 확인시켜 준다. 고용률의 경우 5개월간의 변화 방향 중에서 3개를 맞추어 준수한 예측률을 보여주었으나, 경제활동참가율은 8월과 10월에 대해서 변화 방향을 정확하게 파악하였다.

〈표 3-6〉 예측 모형 III의 변화 방향 예측 결과

	실제값			예측치		
	경제활동 참가율	실업률	고용률	경제활동 참가율	실업률	고용률
2020년 7월	−	−	+	0	−	+
2020년 8월	−	−	−	−	+	−
2020년 9월	+	+	−	−	+	+
2020년 10월	+	+	+	+	−	+
2020년 11월	+	−	+	−	+	−
정확도				2/5	2/5	3/5

자료 : 저자가 설계한 모형의 결과.

제4절 예측 모형들 간의 비교

앞서 우리는 세 가지 모형을 가지고 노동시장의 주요 변수들인 경제활동참가율, 실업률, 고용률에 대한 변수들의 움직임을 예측하고 정확성을 평가하였다. 평균 제곱 오차로 파악한 모형의 정확성은 경제활동참가율의 경우 모형 Ⅰ > 모형 Ⅲ > 모형 Ⅱ의 순서였으며, 실업률과 고용률의 경우 모형 Ⅰ > 모형 Ⅱ > 모형 Ⅲ 순서였다. 반면, 지표들의 증감 여부를 놓고 평가한 결과에서는 경제활동참가율의 경우 모형 Ⅱ > 모형 Ⅰ = 모형 Ⅲ, 실업률에 있어서는 모형 Ⅰ = 모형 Ⅱ = 모형 Ⅲ, 고용률은 모형 Ⅰ > 모형 Ⅱ = 모형 Ⅲ의 순서로 나타났다.

두 결과를 종합한다면 실업률과 고용률에 대해서는 모형 Ⅰ의 예측력이 가장 좋은 것으로 추정되고, 경제활동참가율의 경우에는 어떤 특정한 모형이 나은 예측력을 보여줬다고 말하기 힘들다. 다만 전반적인 예측력에 있어서는 모형 Ⅰ이 가장 괜찮은 반면, 모형 Ⅲ이 가장 떨어지는 예측력을 보여준 점이 특기할 만하다.

이러한 결과를 통해 우리는 다음과 같은 결론을 내려볼 수 있다. 첫째, 노동시장의 경우 내생성과 계절성이 매우 큰 것으로 보인다는 점이다. 특히 고용률이 그러한데, 만 15세 이상 64세 이하의 생산가능 인구 중에서 취업자의 비율이 얼마나 되는지를 측정하는 고용률은 계산을 위해 사용하는 분모와 분자의 두 지표 모두가 전체 생산가능 인구나 취업자 전체를 대상으로 하는 것이어서 수치가 비교적 안정된 데다가 변동폭이 매우 크지 않아서로 보인다. 경제활동참가율의 경우 분모의 숫자는 전체 생산가능 인구로 고용률과 동일하지만 분자의 숫자가 취업자와 실업자의 합인데, 실업자의 숫자가 노동시장 외적인 요인에 의해서 크게 변화할 수 있어서 고용률과 비교하면 상대적으로 지표의 안정성이 낮아서 모형들의 예측력이 조금씩 낮게 나타난 것으로 보인다.

둘째, 실업률의 경우 모든 모형에서 예측이 가장 힘든 지표로 나타났는데, 이는 실업률을 계산하는 분모에 들어가는 경제활동인구부터가 여러

경제적 요인에 의해 영향을 받으며, 또한 생산가능 인구에 비해 작기 때문에 실업자 수의 변화가 비교적 큰 비율의 변화로 나타나는 것도 시계열상의 안정성을 떨어뜨려 정확한 예측을 힘들게 하는 것으로 추측된다.

실업자의 숫자도 여러 요인에 의해서 영향을 받는 상대적으로 불안정한 변수이다. 특히 전체 실업자의 수 자체가 경기가 어떠한 상황이건 크게 늘거나 줄어드는 것이 모두 가능하다. 경기가 악화되어 고용원의 숫자가 줄어들었다 하더라도 만일 장기적인 경기 변동을 비관적으로 보고 기존 실업자들의 일부가 일정 기간 구직활동을 하지 않는 경우에는 설령 취업 상태가 아니라고 해도 이들이 비경제활동인구로 간주되어 실업 통계에서 제외되고, 이로 인해 실업자의 수가 줄어들 수도 있다. 반면, 불경기로 인해 취업자의 일부가 실직 상태에 빠져 바로 실업자로 넘어오는 경우에는 즉각적으로 실업자 수가 크게 증가할 수 있다. 반대로 호경기인 경우에는 기존에 구직활동을 하지 않던 비경제활동인구 일부가 바로 구직활동을 개시하여 실업자로 편입된다면 오히려 경기 여건이 개선되는데도 불구하고 실업률이 증가할 수 있다. 반대로 경제활동이 활발해지면서 실직자 일부가 바로 직장을 찾아서 취업자로 전환되는 경우에는 단기간에 실업자의 숫자가 줄어들 수 있다. 즉, 호경기에도 실업자 수 자체는 단기적으로 늘어날 수도 있고 줄어들 수도 있으며, 불경기에도 양방향으로의 실업률 증감이 모두 가능하기 때문에 수치의 변화폭이 경제활동참가율이나 고용률과 비교하여 상대적으로 크게 나타났다.

셋째, 모형 I의 전반적인 예측력이 가장 좋으며 모형 III의 예측력이 가장 낮은 것으로 나타난 결과를 통해 기계학습에서 노이즈를 제거하거나 통제하는 것이 모형의 예측력을 높이는 데 중요하다는 점을 보여준다. 일반적으로 심층학습이나 기계학습에서는 데이터가 많은 것이 예측력을 높이거나 모형의 성능을 개선하는 경우가 많은데, 만일 많은 데이터들이 노이즈로 작용한다면 drop-out rate을 설정한다 하여도 상대적으로 예측 지표와 무관한 변수들이 drop-out 되지 않고 남아서 모형이 과적합되는 문제를 만드는 것으로 보인다. 모형이 과적합하는 경우에는 다양한 독립변수를 활용하여 종속변수의 움직임을 비교적 정확하게 설명하는 데에는 유용하지만, 예측에 있어서는 종속변수와 무관하거나 관계가 거의 없는

변수들에게까지 설명력을 부여하여 오히려 변수의 예측을 부정확하게 만들 수 있다.

넷째, 모형 I이 모형 II보다 전반적인 예측력이 나았다는 점에서 노동시장은 여타 실물 경제 시장과 깊게 연결되어 있지 않은 시장으로 보이며, 특히 내생성이 큰 것으로 보인다. 이는 한편으로는 노동시장이 경직적이어서 외부 실물 경기의 변동과 비교하여 둔감하게 움직일 가능성이나 시차를 두고 뒤늦게 상대적으로 작은 폭으로만 변화할 가능성도 아울러 시사한다. 또한 모형 III의 예측력이 전반적으로 낮은 점을 통해 여론이나 사람들의 심리가 노동 수요나 노동 공급에 미치는 영향 역시 제한적인 것으로 보인다.

노동 수요가 여론에 의한 영향을 덜 받는 것은 노동시장의 경직성으로 인해 기업들이 경험으로 체득하여 시장의 경직성에 일정하게 대응한 것으로 해석할 수도 있다. 한편 노동 공급 역시 여론이나 심리의 영향을 많이 받지 않는다는 것은 경제 주체들이 노동 공급에 관한 의사결정을 할 때 여론이나 언론을 고려하기보다는 실물 변수나 노동시장 자체의 상황에 근거하여 판단을 내린다고 해석할 수 있다. 즉, 노동시장에서 수요자와 공급자 모두 의사결정 과정에서 노동시장의 당시 상황을 가장 중요한 변수로 고려하며 실물 경제 변수를 일부 고려할 수 있으나 크게 고려치 않으며, 여론이나 언론의 동향은 별다른 고려 대상이 아닌 것으로 보인다. 이는 노동시장에 영향을 미칠 만한 실물 변수들의 영향이 이미 직전 몇 개월간의 노동시장의 움직임에 반영되어 있기 때문일 수 있다.

이러한 결론을 경제 현상에 대한 여론이나 언론의 영향 전반에 대한 평가로 확대 해석하는 것은 경계해야 한다. 이는 경제 현상 자체가 여러 실물 변수로 구성되어 있는데, 본 연구에서는 실물 변수와 신문 기사 제목 간의 상관관계를 살펴보거나 방향의 일치성을 연구한 것은 아니기 때문이다. 결과에서 나타난 모형 간의 예측력 차이는 노동시장이 여타 재화나 서비스 시장과 비교하여 가지는 특수성에서 기인할 수도 있으며, 혹은 모형에서 사용한 데이터의 종류와 개수, 그리고 비정형 데이터를 어떻게 처리하였는지에 따라서 영향을 받을 수 있다. 모형의 설계를 바꾸었을 때 다른 결과가 나타날 수 있는 만큼, 여론이 노동시장에 미치는 영향이 제

한적일 수 있다는 결론은 가능하여도 영향이 없다고 결론 내리는 것은 과도한 해석이라 할 수 있다.

개별 변수에 대해서 모형의 예측 결과를 분석할 필요도 있다. 우선 경제활동참가율의 경우, 2020년 6월에는 63.2%로 2020년 11월까지 중 가장 높은 수치를 보였다가 2020년 8월 62.4%로 최저치를 찍은 이후 2020년 11월 62.8%로 서서히 회복하였다. 특히 2020년 8월에 전월과 비교하여 0.7%p라는 비교적 큰 수치로 떨어졌는데, 이는 1분기 이후 진정세에 접어든 COVID-19에 의해 경제활동이 활발해지고 수출이 회복세에 접어듦에 따라 고용 지표도 개선되었으나, 8월에 광화문 집회발 COVID-19 재확산과 그에 따른 거리두기 단계 격상의 영향으로 경제활동이 전반적으로 크게 위축되었기 때문으로 보인다. 모든 모형에서 2020년 8월에는 경제활동참가율이 떨어질 것으로 전망했는데, 모형 I의 경우 0.5%p, 모형 II의 경우 0.4%p, 모형 III의 경우에는 0.5%p 하락할 것으로 예측했다. COVID-19의 경우 통계 집계상으로는 관찰되지 않는 경제 외적인 요인이므로 이러한 외생 변수에 대한 영향까지 고려할 수 있는 모형 III에서 가장 나은 예측력을 보여야 하지만 모형 I과 모형 III 간에 하락률에서는 차이가 없었다. 오히려 세 지표 모두 8월의 하락에 대해서 이를 추세적인 것이라 보고 9월 전망치에서 큰 폭의 하락을 반영하였는데, 실제 9월에는 8월과 비교하여 경제활동참가율이 0.1%p 올랐지만, 모형 I에서는 0.6%p, 모형 II에서는 0.7%p, 모형 III에서는 0.6%p 하락하는 것으로 예측했다.

이러한 결과는 8월에 있었던 경제활동참가율의 하락이 계절적 요인에 의해 발생한 것이 아닐 가능성을 시사한다. 만일 8월의 하락이 추세적인 것이었거나 혹은 계절적인 요인에 의해서 매년 반복되었다면 모형 I에서는 상대적으로 정확하게 이러한 계절적 변동을 예측하고 반영했을 것이나 실제로는 다른 두 모형과 마찬가지로 모형 I에서도 1개월 늦게 8월의 큰 폭의 경제활동참가율의 하락을 반영하였다. 그러나 노동시장 이외의 외생적인 요인에 의한 변화임에도 모형 III에서도 변화의 크기가 다른 모형과 비교하여 크게 다르지 않은 점은 다음의 세 가지 가능성을 시사한다. 첫째는 COVID-19이 처음 나타났던 2020년 2월 이후 COVID-19에 대한 방역이 비교적 성공적으로 이루어짐에 따라 COVID-19에 대한 언

급이 지속되고 있음에도 2분기부터 실물 경제와 고용이 회복세를 보이면서 시스템이 COVID-19이 고용 지표를 크게 악화시키지 않는다고 인식했거나 혹은 COVID-19이 고용 지표와 무관한 사회경제적 활동이라고 인식했을 가능성이 있다. 둘째로, 8월 전반기의 심각하지 않았던 COVID-19 상황과 후반기의 급증한 COVID-19 확진자 수 및 사회적 거리두기 격상에 따른 8월 전체의 여론 동향을 분석하였음에도 8월 전반기 동안의 낮은 언급량에 의해서 8월 전 기간의 평균 지표가 아주 크게 나쁘지는 않는 것으로 전망했을 가능성이 있다. 즉, 월 초반의 COVID-19이 완화되는 방향의 기사들과 월 후반의 COVID-19이 심각해지는 방향의 기사들이 혼재됨에 따라 8월 전반기는 금년도 상반기와 마찬가지로 고용 지표가 개선될 것으로 예측하였으나, 8월 후반기의 COVID-19 언급량 급증과 사회적 거리두기 단계 격상에 따라 경제활동참가율의 하락을 예상하여 해당 월 전반기의 경제활동참가율 향상에 의해서 후반기 하락분이 일부 상쇄되었을 수 있다. 마지막으로 COVID-19에 대한 본격적인 언급이 시작된 것이 2020년 초이기 때문에 아직 COVID-19에 대한 학습이 충분히 이루어지지 않았을 가능성이 있다. COVID-19 언급량과 고용 지표 간의 변화를 학습할 수 있는 기간이 2020년 2월부터라고 해도 고작 6개월 남짓이기 때문에 COVID-19과 고용 지표 간의 관계를 명확하게 파악하기에는 짧은 기간이어서 아직은 예측력이 크게 개선되지 않았을 수 있다. 이 경우 이후 분석 가능한 기간이 증가하면 모형 III의 예측력이 크게 올라갈 수 있는 가능성을 시사한다.

실업률에 대한 모형의 예측 결과를 분석하면, 실업률의 경우에는 8월에 0.9%p 크게 하락하였다가 다시 9월에 0.5%p 상승한 후 대체로 비슷한 값을 유지하고 있다. 이에 대해서 모형 I은 8월에 0.2%p, 9월에는 0.9%p 하락하는 것으로 예측했고, 모형 II는 각각 0.3%p와 0.8%p 하락, 모형 III은 변화가 없다가 1% 하락하는 것으로 예측했다. 실업률이 급격히 변화하는 시기에 대해서 모형 II가 변화의 크기 측면에서 가장 나은 예측력을 보여준 것은 실업률을 예측하는 것에 있어서 다른 두 변수와는 상이한 특성이 존재할 가능성을 시사한다. 첫째, 실업률에 영향을 주는 경제적 충격은 마찬가지로 실물 경제에도 영향을 주는 변수일 가능성이 높다. 단순히 평

균 제곱 오차만을 놓고 보면 모형 II 내에서 실업률의 평균 제곱 오차가 가장 크게 나타났으며 모형 I의 실업률 예측 평균 제곱 오차가 가장 작지만, 7월에서 9월까지의 실업률 변동에 대해서는 모형 II의 예측이 가장 근사했다. 따라서 이 시기에 발생했던 경제 내외적인 충격은 실물 경제에도 유사한 방향으로 영향을 주어서 실업률을 변화시킨 것으로 보이며, 그 결과 해당 시기 모형 II의 실업률에 대한 예측력이 다른 두 모형에 비해 낫게 나타난 것으로 보인다.

둘째, 해당 기간 실업률 변수가 경제활동참가율이나 고용률에 비해서 불안정적한 추세를 보였는데, 이를 통해 변수의 안정성이 떨어지는 지표에 대해서는 기계학습의 예측력이 떨어진다는 점이다. 경제활동참가율 역시 8월에 지표가 급격히 떨어졌다가 9월에 상승하였으나 경제활동참가율 지표의 평균 제곱 오차는 모든 모형에서 0.099~0.12 사이의 낮은 값을 나타낸 반면, 실업률의 경우 평균 제곱 오차가 0.29~0.31 사이로 나와서 단순히 평균 제곱 오차만 보면 약 3배 정도의 차이가 나는 것을 알 수 있다. 실제 변수의 변화에서의 절대값은 경제활동참가율이 실업률에 비해서 높지만, 실업률이 경제활동참가율에 비해 워낙 낮은 값을 가지고 있기 때문에 변화율에 있어서는 실업률이 더 높게 나타났다. 따라서 예측에 있어서 중요한 변수의 안정성은 변화율에 있지 변화하는 크기의 절대값이 아니라는 점을 확인할 수 있다. 이러한 결과는 앞서 살펴본 Vargas et al.(2018)과 Cerchiello et al.(2018)의 결론을 비교하면 기계학습 모형은 단기 변동성이 큰 수치를 예측하는 데에는 적합하지 않았으나 비교적 안정된 움직임을 보이는 중장기 변수에 대한 예측력에서는 뚜렷한 성능 개선을 보였기 때문이다.

고용률 변수는 예측 대상 기간 내내 가장 안정적인 움직임을 보이는 변수였다. 해당 기간 고용률 변수의 한 가지 특징은 다른 두 지표값이 8월에 크게 하락하였다가 9월에 반등한 반면, 고용률 지표는 큰 하락이나 상승이 없었다는 점이다. 그러나 이것이 고용률 지표값이 세 값 중 가장 안정적이라는 의미는 아닌 것이, 만일 2001년부터 2019년까지 과거 19년간 존재했던 계절적 요인을 반영한다면 예측 대상 기간인 5개월간 고용률이 크게 상승하거나 감소해야 하는 지표였다면 오히려 8월에 발생한

외부적 요인에 의해 고용률 값이 크게 변화하지 않은 것으로 보는 것이 타당하기 때문이다. 이는 세 모형의 8월 예측값을 통해 확인해 볼 수 있다.

8월의 고용률에 대해서 모형 I은 전월 대비 0.4%p 하락하는 값을 예상했으나 모형 II는 0.3%p 하락하는 것으로, 그리고 모형 III은 무려 0.6%p 감소한다고 전망하였다. 이를 통해 계절적 요인 및 내생적 요인을 고려하면 8월에는 고용률이 실제 0.1%p보다 더 크게 감소해야 하지만, 외생적으로 발생한 경제적 충격이 아직 실물 부문에 완전히 전달되지 않은 데다가 노동시장이 상대적으로 경직적이다 보니 실제 실업률의 감소는 실물경제까지 고려하면 덜 감소하였으며, COVID-19에 의한 여론 효과까지 고려한 모형 III에서는 고용률이 더욱 크게 감소할 것으로 전망되지만 이러한 예측 결과를 통해 해석하자면, 여론에서 COVID-19의 효과가 실제보다 과장되어 반영되었을 가능성을 발견하였다. 한편 세 모형 모두에서 2020년 11월에는 실제와는 달리 고용률이 크게 감소하였을 것으로 예측되었는데, 이는 지난 19년간 나타났던 계절적 요인이 영향을 미친 것으로 보이며, 그러나 COVID-19에서의 회복으로 인해 수출이 호조를 보이는 등의 영향으로 지난 19년과 다른 패턴의 결과가 나온 것으로 보인다.

고용률에 대한 세 모형의 예측력이 경제활동참가율보다는 다소 낮은 것으로 기록되었으나 실업률에 비해서는 현저히 높은 것으로 나타났다. 경제활동참가율과 평균 제곱 오차는 다소 높게 나타났으나 큰 차이가 나지는 않았으며, 세 모형 모두에서 변화의 방향을 예측한 데 있어서는 세 가지 지표들 중 가장 나은 모습을 보였다. 따라서 변수의 안정성이 방향성을 정확하게 예상하거나 예측력 자체를 높이는 데 있어서 주요한 변수라는 점을 확인할 수 있다.

한편 본 모형은 기계학습을 통해 경제활동참가율, 실업률, 고용률이라는 세 가지 변수의 예측력과 설명력을 동시에 살펴볼 수 있다. 5개월 각각에 대해서 전월까지의 모든 정형 및 비정형 데이터를 활용하여 변수를 예측케 했다는 점에서는 예측력을 평가한 모형이라 할 수 있으며, 한편으로 우리는 각 월에 대해서 지난 월까지의 변수값 모두를 알고 있을 뿐만 아니라 당월의 예상 변수값에 대해서도 실제 관측치를 알고 있기 때문에 전월까지의 변수들이 당월의 고용 지표들을 얼마나 잘 설명하는지 설명

력을 살펴보는 것이기도 하다. 본 연구에서 이렇게 설명력과 예측력을 동시에 살펴보는 방식으로 모형을 설계한 것은 두 가지 이유에서이다.

첫 번째 이유는 데이터의 한계이다. 본 연구에서 사용한 정형 데이터와 비정형 데이터들은 모두 각각 출간 및 공개되는 시점이 다르다. 따라서 일부 데이터는 비교적 이른 시기에 입수 가능한 반면 일부 데이터는 한참의 시간이 경과되어야 입수 가능하다. 따라서 가능한 한 많은 양의 정형 데이터를 분석에 포함시키기 위해서 분석 시점 대비 일정한 시차를 둔 과거의 자료를 분석하였기 때문에 예측력과 설명력을 동시에 살펴보는 방식으로 진행되었다.

다른 이유는 모형의 성능을 평가하기 위해서이다. 모형을 통해 예측하는 것 그 자체에만 의의를 두는 것이 아니라 분석에 사용되는 변수들을 달리하면서 세 가지 모형의 성능을 비교하고 이를 통해 각 모형이나 데이터의 성격, 그리고 노동시장 지표들의 분석 가능성을 살펴보려 했기 때문에, 모형으로 하여금 노동시장 변수들을 예측하게 하였음에도 동시에 노동시장 변수에 대한 예측값을 실제값과 대조하여 얼마나 모형이 정확한지를 살펴보는 것도 겸하는 방식으로 설계하였다.

제5절 소 결

본 장에서는 실제 기계학습 모형을 설계한 후, 각 모형마다 심층학습에 사용하는 데이터의 범위를 달리하여 노동시장의 주요 변수들인 경제활동참가율, 실업률, 고용률을 예측토록 하였다. 첫 번째 모형은 노동시장의 주요 변수들의 시계열 자료만을 활용하여 노동시장의 움직임을 계절성과 내생성, 그리고 노동시장의 자체적인 장단기 추세만을 이용해 예측토록 하였다. 두 번째 모형은 각종 실물 경제 변수와 경기 현황 및 전망, 수출입 등에 대한 다양한 정형 데이터를 포함하여 예측하였다. 마지막으로 세 번째 모형에서는 두 번째 모형에서 사용한 모든 정형 데이터에 신문 기사 제목이라는 비정형 데이터까지 포함하여 노동시장 변수들을 예측토록 함

으로써 노동시장의 주요 변수들이 여론이나 심리의 영향을 받는 경제 변수인지를 알아보도록 하였다. 모형의 예측 결과는 다음과 같았다.

경제활동참가율의 경우 평균 제곱 오차로 측정한 모형의 예측 결과는 노동시장 변수만을 고려한 모형 I의 예측력이 가장 좋고 모형 II의 예측력이 가장 낮았으나, 경제활동참가율의 월별 증감을 얼마나 정확하게 예측했는지까지 고려한다면 어떤 한 모형이 우월하다고 말하기 어려운 결과가 나타났다. 실업률의 경우 모형 I이 평균 제곱 오차나 변수의 방향성 모두에서 가장 나은 것으로 나타났으며, 고용률 역시 마찬가지로 모형 I의 예측력이 모든 지표에서 가장 좋은 것으로 분석되었다.

이는 다음과 같은 사실을 함의하는 것으로 보인다. 첫째로 노동시장은 내생성이 강한 시장이라는 점과 아울러, 여타의 실물 경제에 즉각적으로 반응하지 않는 일정한 정도의 경직성을 가지는 것으로 보인다. 이로 인해서 다수의 실물 경제 변수를 포함한 모형 II의 예측력이 전반적으로 모형 I에 비해서 좋지 않게 나온 것으로 보인다.

반면 노동시장의 경우 사람들의 심리나 여론 등에 별다른 영향을 받지 않으며, 상대적으로 실물 경제의 움직임이나 노동시장 자체의 상황에 보다 민감하게 반응하는 것으로 보인다. 경제 주체들이 노동시장에서 노동공급과 관련된 결정을 할 때 실물 경제의 움직임이나 현재 노동시장에서 취업 확률이 얼마나 되며 최근 어떠한 움직임을 보이는지는 고려하는 것으로 보이는 반면, 사회적 이슈나 최근의 여론의 동향이 직접적으로 노동공급 함수에 영향을 주지는 않거나 혹은 그 영향이 제한적인 것으로 결과가 나타났다. 이는 모형 III이 실업률과 고용률에 있어서는 모든 지표에서 예측력이 가장 좋지 않았으며, 경제활동참가율에 있어서도 변수의 움직임에 대해서는 예측력이 좋지 않았고 다만 평균 제곱 오차에 있어서만 모형 II보다 나은, 좋지 않은 결과를 보였다는 점에서 확인할 수 있다.

이러한 결과가 나타나게 된 한 가지 이유는 모형 II나 모형 III에서 사용한 변수들 중 일부가 노이즈처럼 작용하여 예측력을 떨어뜨렸기 때문으로 파악된다. 일반적으로 기계학습의 경우에는 데이터의 크기가 커지면 커질수록 모형의 설명력과 예측력이 올라가지만, 여기서 사용한 각종 데이터들은 노동시장과 무관한 변수들까지 포함시키다보니, 모형이 일부

에서 과적합하거나 혹은 노동시장과 무관한 변수들에까지 설명력을 부여하여 오히려 모형 전반의 예측력이 떨어지게 나온 것으로 보인다.

마지막으로, 그러나 본 장에서의 분석 결과가 노동시장이 사람들의 심리나 여론에 대해서 전혀 영향을 받지 않는다고 해석하는 것은 경계해야 한다는 점이다. 이는 본 연구에서 지난 20년간의 모든 신문 기사 제목을 포함시키다보니 노동시장과 무관한 여러 신문 기사나 사건들이 포함됨에 따라 노이즈로 작용하였을 수 있다. 노동시장과 직접적으로 연관된 기사들만 포함시킬 경우 예측력이 올라갈 수 있으며 또한 자연어를 처리하는 학습 방법을 바꿀 경우에도 역시 모형의 예측력은 바뀔 수 있다.

제4장
결 론

본 연구는 노동시장의 주요 변수들인 경제활동참가율과 고용률, 실업률을 기계학습과 심층학습을 통해 예측해 보았다. 이때 예측을 위한 변수로 노동시장의 각종 변수들만을 포함한 모형 I과 노동시장 변수를 포함하여 모든 가용한 통계 자료를 사용한 모형 II, 그리고 모형 II에 지난 20년간의 모든 신문 기사 제목을 포함한 모형 III 등 세 가지 모델을 설정한 후 이들의 성능을 비교하였다.

이를 통해 노동시장은 내생성이 강하여 모형 I이 전반적으로 가장 나은 예측력을 가진 것으로 나타났으며, 여론 및 사람들의 심리를 대변하는 신문 기사 제목까지 사용한 모형 III의 예측력이 가장 낮은 것으로 나타났다.

그러나 이러한 예측 결과를 일반화하여 노동시장의 주요 변수들을 노동시장의 내적 요인들로만 설명하는 것으로 충분하다고 생각하는 것은 성급한 결론이다. 앞서 제2장에서 언급하였으나, 본 연구는 일반적인 빅데이터 분석에서 지적하는 한계점을 가지고 있다. 보통 기계학습에서 학습 데이터를 이용하여 모형을 최적화할 때 과적합을 방지하기 위해서는 최소 10만여 개 이상의 학습 데이터가 필요하다고 보고 있으나, 본 연구에서는 경제 변수를 예측한다는 연구 조건의 한계상 지난 20여 년간의 약 240여 개의 학습 데이터만을 사용하였다. 따라서 과적합의 가능성을 피할 수 없으며, 시간이 흐르면서 학습 가능한 데이터가 누적될 경우 연구의

결과는 언제든지 바뀔 수 있다. 그러나 20여 년간의 데이터를 바탕으로 분석한 결과에 따르면 다음과 같은 결론을 얻을 수 있다.

첫째, 실물 경제의 주요 변수들을 예측하는 데 있어 여론이나 언론의 영향은 생각만큼 크지 않을 가능성이 높다. 이는 사람들이 경제 행위에 대해서 의사결정을 할 때, 적어도 노동 공급에 있어서는 실물 경제의 움직임에 보다 민감하게 반응하기 때문인 것으로 보인다. 이는 모형 II가 실업률과 경제활동참가율이 급격하게 변했던 2020년 8월을 전후하여 예측력에 있어서 다른 모형들과 비교하여 괜찮은 성능을 보였던 데서 확인 가능하다.

이러한 결과가 나오는 이유 중 하나로는, 신문의 경우 실물 경제에 영향을 미치는 주요한 사건이라 하더라도 논쟁적이지 않거나 사람들의 주목을 받지 못하는 경우에는 많이 다루지 않는다는 특성이 있기 때문이다. 따라서 논쟁적이거나 주목을 받는 주제가 상대적으로 과대 대표되어 실제 경제 변수들에 대한 예측력을 방해하는 것으로 작용할 수 있다.

또한 본 연구에서는 개별 신문 기사가 모두 동일한 가중치를 가지고 여론을 대변하거나 경제 주체들의 심리에 영향을 미쳤다고 가정하고 분석하였으나, 실제로는 주제나 기사별로 사람들이 주목하는 정도가 다를 수 있다. 이를 보정하기 위한 한 가지 방안으로 각 기사별로 사람들이 얼마나 읽었는지를 알려주는 클릭수 자료를 추가하는 것을 고려할 수 있다. 하지만 신문 기사는 각 신문사의 웹페이지 뿐만 아니라 다양한 포털 사이트에도 게시되고 있기 때문에 정확한 클릭수를 집계하는 것이 불가능하다는 문제점이 있다. 또한 클릭수를 제공하지 않는 포털 사이트나 신문사의 경우 해당 자료가 원천적으로 수집 불가능하여 활용할 수 없기 때문에 본 연구에서는 이를 포함시킬 수 없었다. 하지만 추후 다른 연구에서 이를 반영하여 기사별로 보다 정확한 가중치를 부여할 수 있는 방안을 찾는다면 신문 기사를 비롯한 비정형 데이터를 포함한 모형의 예측력이 높아질 수 있을 것이다.

둘째로, 실물 경제나 노동시장이 여론이나 심리에 의해서 영향을 받지만 그것이 신문 기사보다는 다른 비정형 데이터에 의해서 대변될 가능성도 있다. 사회관계망 서비스(Social Network Service)나 인터넷 커뮤니티

의 활동이 이러한 여론을 대변하는 변수라면 비정형 데이터의 대상을 바꿈으로써 모형의 예측력을 개선하거나 심리의 영향을 보다 잘 반영할 수 있을 것이다. 그러나 사회관계망 서비스나 인터넷 커뮤니티의 글들은 시간이 흐름에 따라 계정이 삭제되거나 글이 변경되거나 지워지는 문제점이 있다. [그림 2-1]에서 살펴본 바와 같이 신문 기사에서도 시간이 흐름에 따라서 신문 기사가 삭제되어 과거의 기사의 양이 최근보다 적은 문제점이 있었는데 사회관계망 서비스나 인터넷 커뮤니티에서는 이러한 현상이 훨씬 심하다. 따라서 다른 형태의 비정형 데이터를 활용하고자 한다면 최근의 수집량이 더 많은 문제를 보정하기 위한 방안을 고려하며, 수집된 과거의 데이터들에서 잔존한 데이터들이 당시의 여론을 정확해 대변하는지를 검증한 후 사용해야 할 것이다.

셋째, 경제 변수를 예측할 때에는 노이즈를 걸러내기 위해서 모형에 사용할 경제 변수들을 고르는 데 있어서 신중해야 할 필요가 있다는 점을 확인할 수 있었다. 일반적으로 기계학습은 사용하는 데이터의 양이 많아지면 많아질수록 예측력이나 설명력이 나아지지만, 만일 다수의 데이터가 예측하고자 하는 변수나 설명하고자 하는 종속변수와 무관하다면 이들에 의해서 오히려 예측력이 저하될 수 있다. 따라서 경제 현상을 설명하거나 예측하는 모형을 설계할 때 변수를 거르지 않고 가용한 모든 자료를 넣는 것이 기계학습이나 심층학습의 취지에도 맞고 모형의 설명력도 높일 수는 있으나, 그 자체가 예측력을 담보하지 않을 뿐만 아니라 오히려 예측력을 저하시킬 수도 있으므로 사전에 모형에서 사용할 변수를 취사선택하는 것이 나을 수 있다. 만일 이러한 작업에 연구자의 임의성을 최대한 배제시키거나 혹은 경제적인 직관을 찾기보다는 설명력이 가장 높은 모형을 구축하는 것만 고려한다면, 라쏘나 릿지 등의 방법을 통하여 모형에 사용할 변수를 우선 탐색한 후, 이러한 변수를 바탕으로 기계학습 및 심층학습을 수행하는 것도 일책일 수 있다.

넷째, 노동시장 변수를 예측하는 데 있어 계절성과 내생성은 고려해야 할 최우선 사항이라는 점이며, 이것이 다른 경제 변수들에서도 그대로 적용될 가능성이 높다. 최근 기계학습을 이용하여 여러 경제 변수들의 움직임을 예측하거나 혹은 전망하려는 시도가 있는데, 이때 모형이 시계열 자

료를 분석하게 하는 방법에는 여러 가지가 있다. 이때 단순히 변수의 추세만을 따라가게 하는 방식보다는 계절성을 넣어서 연 단위, 혹은 일정한 기간마다 반복되는 추세를 통제하는 것이 중요하며, 또한 예측하고자 하는 변수와 직접적으로 관련된 변수들만을 통해 예측하는 것이 가장 쉽고 괜찮은 예측력을 얻을 수 있는 방법으로 보인다.

다섯째, 본 연구에서 세 가지 모형 모두 방향성을 예측하는 데 있어 그다지 좋은 성과를 보여주지 못했다는 점을 통해 학습 데이터의 누적이 중요하다는 것을 확인할 수 있었다. 따라서 만일 기계학습을 통해 예측하고자 하는 지표가 있다면, 최대한 시계열이 길며, 관련 데이터 역시 긴 시간 동안 제공되는 것을 선택하는 것이 예측의 정확성을 높일 수 있는 길로 보인다.

본 연구는 기계학습을 이용하여 경제 변수를 예측하여 보았다. 경제 변수를 기존의 회귀 분석이나 라쏘 및 릿지 등의 방법이 아니라 순수한 기계학습으로 진행한 것은 향후 기계학습 및 심층학습을 경제 각 분야에 적용하고 여러 방법론과 연구를 진일보하게끔 하는 촉매가 될 수 있다. 다만, 기계학습이 경제학 연구에서 널리 쓰이지 않는 중대한 이유가 있음도 연구 마지막에 언급할 필요가 있다. 본 연구에서도 연구 결과를 해석할 때 단정적인 해석을 최대한 지양하고 추정하는 형식으로 기술하였다. 이는 기계학습이 모형의 결과를 내놓기는 하지만 왜 이러한 결과가 나왔는지에 대해서는 전혀 설명을 제공하지 않기 때문이다. 만일 우리가 관심 있는 사안이 왜 모형이 이러한 결과를 냈는지, 혹은 관심이 있는 경제 변수 간에 어떠한 관계가 있는지를 직접 알아보고 싶은 것이라면, 기계학습은 적합한 분석 방법이 아니며, 오히려 기존의 방법들이 훨씬 직관적이면서도 명쾌하다. 따라서 모든 경제 분석에 기계학습을 적용하는 것은 자칫 정책적 함의나 실물 경제에의 적용 가능성을 배제한 채 분석 자체만을 위한 분석이 될 가능성이 높다. 그러므로 연구자는 연구 설계 단계에서 진행하고자 하는 연구가 기계학습에 적합한 연구인지 충분히 검토하고 실제 연구 진행 과정에서도 다양한 방향으로 분석을 진행함으로써 학습 결과에 대한 설명을 제공하지 않는 기계학습의 단점을 충분히 극복할 수 있는 방안을 마련해야 할 것이다.

참고문헌

김수현 · 손욱(2020), 「북한 경제연구로 분석한 경제정책 변화: 텍스트 마이닝 접근법」, BOK경제연구 2020-6호.

방형준 · 손연정 · 노세리(2019), 『데이터 산업의 노동시장 분석』, 한국노동연구원.

정한웅(2016), 「딥러닝 알고리즘에 기반한 기업부도 예측」, 한양대학교 대학원 석사학위 논문.

통계청, 「경제활동인구조사」.

Bernanke, B. S.(1983), "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review* 73(3), pp.257~276.

Cerchiello, P., G. Nicola, S. Rönnqvist, and P. Sarlin(2018), "Deep Learning for Assessing Banks' Distress from News and Numerical Financial Data," SSRN Electronic Journal.

Kim, Soohyon(2020), "Macroeconomic and Financial Market Analyses and Predictions through Deep Learning," BOK 경제연구 2020-29.

Le, Q. V. and T. Mikolov(2014), "Distributed Representations of Sentences and Documents," 32.

Mitchell, Tom(1997), *Machine Learning*, New York : McGraw Hill.

Sheehan, E., C. Meng, N. Jean, M. Tan, M. Burke, S. Ermon, B. Uzkent, and D. Lobell(2019), "Predicting Economic Development using Geolocated Wikipedia Articles," Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.2698~2706.

Soe, A. M.(2014), "Does Past Performance Matter? The Persistence

Scorecard," S&P Dow Jones Indicies, McGraw Hill Financial.

Vargas, M. R., C. E. M. dos Anjos, G. L. G. Bichara, and A. G. Evsukoff (2018), "Deep Learning for Stock Market Prediction Using Technical Indicators and Financial News Articles," Proceedings of the International Joint Conference on Neural Networks, 2018-July (February 2019).

〔부 록〕

본 연구에서 사용한 정형 데이터의 노동시장 변수 외 목록

변수명	추출 대상	출처
국내인구이동통계	전국 및 광역시도별로 총전입인구, 총전출인구, 순이동, 시도내이동-시군구내, 시도내이동-시군구간 전입, 시도내이동-시군구간 전출, 시도간 전입, 시도간전출	통계청,「국내인구이동통계」
국제인구이동통계	내국인 총이동, 외국인 총이동, 내국인 입국자, 외국인 입국자, 내국인 출국자, 외국인 출국자, 내국인 국제순이동, 외국인 국제순이동	통계청,「국제인구이동통계」
내국인-출국연령별	전국 및 광역시도별로 연령별, 승무원	한국관광공사,「한국관광통계」
외래객 - 입국목적별 및 국적별	지역별, 입국목적별	
경기종합지수	선행종합지수, 동행종합지수, 후행종합지수	통계청,「경기종합지수」
경제심리지수	경제심리지수(원계열)	한국은행,「경제심리지수」
기업경기조사 전국실적	산업별 업황실적, 매출실적, 채산성실적, 자금사정, 인력사정	한국은행,「기업경기조사」
기업경기조사 전국전망	산업별 업황전망, 매출전망, 채산성전망, 자금사정 전망, 인력사정 전망	한국은행,「기업경기조사」
생산확산지수	산업별 생산확산지수, 증가업종 수, 보합업종 수, 감소업종 수	통계청,「경기종합지수」
전산업생산지수	산업별 원지수	통계청,「전산업생산지수」
설비투자지수	설비투자 총지수, 설비투자 기계류, 설비투자 운송장비	통계청,「설비투자지수」
내수 수출 광공업 출하지수	산업별 내수출하지수, 수출출하지수	통계청,「광업제조업동향조사」
소재부품산업 동향조사	산업별 수출액 및 수입액	산업통상자원부,「소재부품산업동향조사」

변수명	추출 대상	출처
수요자 기종별 기계수주 불변금액	수요자별 원동기, 특수산업용기계, 금속공작·가공기계, 일반산업용기계, 사무자동처리기계, 통신기계, 전기기계, 도로주행차량, 기타수송용기계	통계청,「기계수주동향조사」
시도 재별 제조업 생산지수	재별 생산지수, 생산자제품 출하지수, 생산자제품 재고지수, 내수출하지수	통계청,「광업제조업동향조사」
제조업 생산능력 및 가동률 지수	제조업 생산능력지수, 제조업 가동률지수	통계청,「광업제조업동향조사」
제조업 재고율	제조업 재고율	통계청,「광업제조업동향조사」
제조업 평균가동률	제조업 평균가동률	통계청,「광업제조업동향조사」
제조업 ICT 생산지수	생산지수, 생산자제품 출하지수, 생산자제품 재고지수	통계청,「광업제조업동향조사」
철강 생산량	조강, 철강재, 형강, 철근, 중후판, 열연강판, 냉연강판, 강관	한국철강협회,「철강통계조사」
품목별 광공업 생산 출하 재고 내수 수출량	품목별 생산량, 출하량, 재고량, 내수량, 수출량	통계청,「광업제조업동향조사」
공항별 통계	유형별 도착 및 출발에 대한 숫자	한국공항공사, 인천국제공항공사,「항공통계」
국내건설수주액	공공과 민간 물량의 건설수주액	대한건설협회,「국내건설수주동향조사」
국내선 노선별 통계	유형별 도착 및 출발에 대한 숫자	한국공항공사, 인천국제공항공사,「항공통계」
국제선 지역별 통계	유형별 도착 및 출발에 대한 숫자	한국공항공사, 인천국제공항공사,「항공통계」
동수별 연면적별 건축착공현황	동수별 및 연면적별 건축착공 면적	국토교통부,「건축허가및착공통계」
동수별 연면적별 건축허가현황	동수별 및 연면적별 건축허가 면적	국토교통부,「건축허가및착공통계」
환승 여객 통계	환승 여객 수와 전체 여객 수	한국공항공사, 인천국제공항공사,「항공통계」
IT산업별/월별 수출 현황	5개 소분류별 수출액	과학기술정보통신부,「ICT수출입통계」

변수명	추출 대상	출처
IT산업별/월별 수입 현황	5개 소분류별 수입액	과학기술정보통신부,「ICT수출입통계」
산업별 서비스업 생산지수	13개 소분류별 생산지수	통계청,「서비스업동향조사」
재별 및 상품군별 판매액지수	내구재별 생산지수	통계청,「서비스업동향조사」
건설공사비지수	건설 전체, 건물건설, 토목건설	한국건설기술연구원,「건설공사비지수」
국내공급물가지수	재화별 물가지수	한국은행,「생산자물가조사」
생산자물가지수	상품, 서비스	한국은행,「생산자물가조사」
생활물가지수	생활물가, 식품, 식품 이외, 전월세, 생활물가 이외, 전월세 포함 생활물가지수	통계청,「소비자물가조사」
소비자물가지수	소비자물가지수	통계청,「소비자물가조사」
수입물가지수 기본분류	총지수, 농림수산품, 광산품, 공산품	한국은행,「수출입물가조사」
수입물가지수 용도별	총지수, 원재료, 중간재, 최종재, 자본재, 소비재, 내구재, 비내구소비재	한국은행,「수출입물가조사」
수출물가지수 기본분류	총지수, 농림수산품, 공산품	한국은행,「수출입물가조사」
신선식품지수	신선식품, 신선어개, 신선채소, 신선과실, 신선식품제외	통계청,「소비자물가조사」
지출목적별 소비자물가지수	12개 지출목적별	통계청,「소비자물가조사」
품목성질별 소비자물가지수	상품, 서비스	통계청,「소비자물가조사」
결제시스템별 통계	총액결제시스템 건수 및 액수, 소액결제시스템 건수 및 액수	한국은행,「지급결제통계」
국제수지	경상수지, 자본수지, 금융계정, 오차 및 누락	한국은행,「국제수지통계」
대출금리 신규취급액 기준 금리	종금사할인어음, 신협일반대출, 새마을금고일반대출, 은행신탁대출	한국은행,「통화금융통계」
수신금리 신규취급액 기준 금리	상호저축은행 1년 정기예금, 신협 1년 정기예탁금, 새마을금고 1년 정기예탁금	한국은행,「통화금융통계」

변수명	추출 대상	출처
수출입총괄	수출건수, 수출금액, 수입건수, 수입금액, 무역수지	관세청, 「무역통계」
신규취급액 기준 금리	저축성수신 금리, 금융채 제외 저축성수신 금리	한국은행, 「통화금융통계」
신용카드	신용카드 이용건수, 이용금액, 발급장수	한국은행, 「지급결제통계」
어음교환 및 부도	1장당 평균금액, 부도업체수	한국은행, 「지급결제통계」
예금은행 예금회전율	예금은행 예금회전율, 요구불예금, 저축성예금	한국은행, 「통화금융통계」
은행공동망	타행환, 현금자동인출기, 전자금융공동망, 직불카드공동망, 자금관리서비스망, 지방은행공동정보망	한국은행, 「지급결제통계」
잔액 기준 금리	저축성 수신 금리, 금융채 제외 저축성 수신 금리	한국은행, 「통화금융통계」
주요 통화금융지표	본원통화, 협의통화, M2, 예금은행 총예금, 예금은행 저축성예금, 예금은행 대출금	한국은행, 「통화금융통계」
지로시스템	은행지로 처리건수 및 처리금액, 일반계좌이체 처리건수 및 처리금액, 자동계좌이체 처리건수 및 처리금액, 대량지급 처리건수 및 처리금액	한국은행, 「지급결제통계」
한국은행 주요계정 말잔	자산합계, 국내자산, 국외자산, 부채합계, 국내부채, 국외부채, 자본	한국은행, 「통화금융통계」
한은금융망	원화자금이체 처리건수 및 처리금액, 외화자금이체 처리건수 및 처리금액	한국은행, 「지급결제통계」
주요 에너지 지표	1차에너지 공급, 최종에너지 소비, 에너지 수입액, 에너지 수입의존도, 석유의존도	에너지경제연구원, 「에너지수급통계」
최종에너지 부문별 소비	산업, 수송, 가정 및 상업, 공공	에너지경제연구원, 「에너지수급통계」

◈ 執筆者

• **방형준**(한국노동연구원 부연구위원)

기계학습을 이용한 노동시장 예측모형 탐색

▪ 발행연월일 | 2020년 12월 24일 인쇄
2020년 12월 30일 발행
▪ 발 행 인 | 배 규 식
▪ 발 행 처 | 한국노동연구원
3 0 1 4 7 세종특별자치시 시청대로 370
세종국책연구단지 경제정책동
☎ 대표 (044) 287-6080 Fax (044) 287-6089
▪ 조판 · 인쇄 | 도서출판 창보 (02) 2272-6997
▪ 등 록 일 자 | 1988년 9월 13일
▪ 등 록 번 호 | 제13-155호

 정가 5,000원

ISBN 979-11-260-0431-7