POSRI 이슈리포트

2018. 12. 20.

빅데이터 시대의 예측 도구, Machine Learning의 올바른 활용법

이성경 책임연구원, 동향분석센터 (sunglee.sk@posri.re.kr)

목차

posco
포스코경영연구원

Executive Summary

○ **예측 모델링은 크게 인과관계에 기반한 이론적(theoretical) 모델링과 상관관계에 기반한 축약형(reduced-form) 접근법으로 분류됨**

- 모델 설계자는 모델을 통한 직관적 해석(good stories)과 예측력 높은 결과(good forecasts) 도출을 목표로 하며, 이에 따라 이론적 적합성뿐 아니라 예측력을 모델의 평가 기준으로 활용

- 문제는 이 두 가지 목표가 종종 상충하거나 경제변수, 샘플 기간, 예측 기간(단기·장기)에 따라 예측 모델 간 우위가 변동

  · 예측력 비교는 실증 분석 결과에 기반한 결과론적 평가로서, 문제의 상황에 맞게 모델러가 예측 모델의 후보군들을 설정해 이들을 비교하는 형태로 이루어짐

  · 이론적 모델은 직관적 해석이 가능하나, 다수의 실증연구 결과에 따르면 실제 예측성과 관련이 높은 변수들만 포함한 간결한(parsimonious) 계량 모델들이 예측성에서는 더 우위에 있다는 결론

○ **빅데이터를 이용한 예측법은 축약형 접근법으로서, 최근 빅데이터를 활용하여 유용한 정보를 추출하는 방법에 대한 연구가 활발해짐**

- 기존 축약형 모델은 빅데이터를 다룰 수 있는 틀을 제공하지 못하며, 경제이론이나 모델 설계자의 직관에 기반하여 몇 개의 설명변수를 선택하여 회귀분석 및 예측

- 반면 AI의 하위 카테고리인 머신러닝은 러닝 알고리즘으로서, 빅데이터에 담긴 정보를 추출·가공하여 경제변수를 예측할 수 있는 틀을 제공

  · 머신러닝은 예측문제 환경에 대한 사전적인 이론적·구조적 지식 없이 주어진 알고리즘을 통해 데이터에 담긴 정보를 추출해내는 귀납적 러닝으로, 다양한 구조의 데이터를 분석 가능

  · 알고리즘을 통해 데이터를 학습하고 해당 경제변수 예측에 관련성이 높은 정보를 추출함으로써 회귀분석이 가능한 규모의 프레임을 제공

○ **러닝 알고리즘에 대한 산업공학·신경물리학·통계학 간의 다양한 학제적 논의가 활발하지만, 실제 활용은 산업 현장 주도로 이루어지고 있는 양상**

- 테크·서비스업체(Google, Facebook, Amazon, Uber, Yelp, Airbnb 등)는 이용자들의 패턴을 학습, 은행·제조업은 의사결정, 상품가격 예측 및 주요 경제변수의 향방 예측 등에 빅데이터를 적극 활용함

- **[AI 활용의 예]** (1) 철강 • 철강원료 • 경제변수 예측을 위한 빅데이터 활용 (2) 매 순간 만 단위로 쌓이는 빅데이터(de-identified)를 활용하여 행동패턴 학습(HR analytics) 및 직원들의 복지향상 • 인사정책 개선 방안 개발

○ **머신러닝 기법은 (1) 빅데이터에 담긴 정보를 활용할 수 있는 틀을 제공하며 (2) 검증 데이터(testing set)에서의 결과는 좋지만, (1) 분석결과의 직관적 해석이 어렵다는 점 (2) 실제 경제예측에서 어떠한 알고리즘도 뚜렷한 우월성을 지니고 있다는 것이 입증되지 않았다는 점("No Free lunch Theorem", Wolpert 1996)이 한계로 지적**

- 최근 BlackRock의 경우 해석력 결여 문제를 지적하며 AI를 활용한 투자 리스크 모델링 접근법을 철회시킴 (Risk.net, 2018.11.12일자)

○ **경제예측성 평가는 현재까지 주어진 데이터를 기반으로 이루어지므로, 현재를 바탕으로 한 결과가 미래에도 유의미할 것인가는 자연스러운 의문**

- 경제변수의 동태는 현존하는 의사결정 프레임에 의해 결정되는 것이고, 현 시점 최적의 알고리즘은 지금까지 채집 가능한 데이터에 기반한 것. 이는 모든 예측 모델링에 해당되는 근본적인 질문으로, "economic forecasting is an art, not a science"는 이러한 한계를 지적

- 특히, 머신러닝 기법은 변수 간의 선제적 가정 없이 빅데이터에 담긴 과거 정보를 알고리즘에 따라 분석하는데, 모델러의 조정 없이 단순히 기계 학습을 할 경우 잘못된 결론이 도출될 수도 있음

  · Amazon은 기존 회사인력의 이력서 정보를 인사채용에 활용하였는데 AI 채용법이 기존 테크산업의 남성지배적인 구조를 고착화한다는 사실을 확인하고 이를 철회하기에 이름 (Reuters, 2018.10.10일자)

○ **예측은 현재 가늠할 수 없는 미래 변수의 비이성적 행동변화나 경제충격을 모델 설계에 반영할 수 없는 본질적 한계가 있으므로, 모델 설계자의 통찰력을 반영한 정성적 평가도 중요**

- 산업 전문가의 직관이 반영된 현실적이고 적시성 있는 알고리즘 설계가 요구되며 시시각각 변하는 정치경제적 상황과 각 산업의 특수성을 고려한 미세조정이 필요. 빅데이터를 활용할 수 있는 프레임이 주는 장점을 활용하되, 질적인 정보나 산업 전문지식을 활용하여 모델을 맥락화하는 방안을 고안

1. 예측을 위한 모델링이란

□ 예측 모델링은 크게 변수들 간의 인과관계에 기반한 이론적 모델 설계와 상관관계에 기반한 축약형 접근법으로 분류[1] 가능 (참고1 참조)

○ 모델 설계자는 모델을 통한 직관적 해석(good stories)과 예측력 높은 결과(good forecasts) 도출을 목표로 하며, 이에 따라 이론적 적합성뿐 아니라 예측력을 모델의 평가 기준으로 활용 (참고2 참조)

- 다수의 실증결과에 따르면 정교한 수학적 가정에 기반한 이론적 모델이나 가능한 많은 변수를 포함하여 규모가 큰 축약형 모델이, 중요한 변수들을 선별하여 회귀분석한(parsimonious) 축약형 모델보다 예측력에 있어 우위에 있지 않다는 결론 (Gü rkaynak, Kısacıkoğlu, and Rossi, 2013)

- 이론적 모델은 직관적•인과적 해석이 가능하여 정책분석에 많이 쓰이고 있고 더 많은 설명변수들을 추가하면 더 많은 정보를 포함할 수 있지만, 항상 더 높은 예측력으로 귀결되지는 않는다는 지적

  ⇒ 즉, 더 복잡하고 포괄적인 모델보다는 예측력 제고와 직접적 연관성이 높은 정보를 선택적으로 이용한 예측 모델링이 좋은 결과를 도출하는 데 유리

□ 활용 가능한 데이터의 양이 풍부해지고 컴퓨터의 연산처리 능력이 향상됨에 따라, 빅데이터를 활용한 축약형 모델링이 각광받기 시작

○ 기존의 축약형 모델링은 회귀분석 방법론의 한계로 인해 포함할 수 있는 설명변수의 개수가 제약됨[2]

- 현실에서는 실제 데이터 생성과정(DGP; data generating process)에 대한 정보가 없어 통용적으로 경제이론이나 모델 설계자의 직관에 기반하여 설명변수들을 선별하게 되는데, 이에 따라 설명변수의 자의적 선택 문제가 지적됨

- 모델 설계자는 변수를 하나씩 늘리거나 줄이면서 여러 조합의 설명변수를 포함하는 다양한 함수 형태를 시도

○ 최근 빅데이터에 담긴 방대한 정보의 양을 처리할 수 있게 해주는 축약형 모델링에 대한 연구들이 활발하게 진행되고 있음

---

[1] 시계열 모델 중 경제이론에 기반한 이론적 모델에는 대표적으로 DSGE(Dynamic Stochastic General Equilibrium)와 Structural VAR(Vector Autoregression)이, 축약형 모델에는 AR(Autoregressive), VAR과 Factor 모델 등이 있음

[2] 샘플의 길이(t)보다 변수의 개수(k)가 훨씬 더 큰 경우(k >> t), 다중공선성(multicollinearity)과 과적합(overfitting)의 문제로 최소자승법(OLS; Ordinary Least Squares)과 같은 기존의 방법론 사용에 제약이 생김("the curse of dimensionality")

- 빅데이터 행렬의 차원을 축소하기 위해 해당 변수 예측에 관련성이 높은 변수들만 선택하는 방법 혹은 수학적 개체인 factor를 추출하는 방법, 혹은 두 가지 방법론을 모두 적용한 접근법들이 많이 사용되어옴

- 머신러닝은 주어진 특정 알고리즘을 이용하여 빅데이터에 담긴 정보를 학습하고 패턴 분석을 가능케 하는 러닝 알고리즘으로 예측 도구로 중요성이 높아지고 있음

○ **최근 인과적 추론에 머신러닝 기법을 활용한 연구가 주목을 받기 시작했는데, 향후 인과적 질문을 해결하는 이론적 모델링에도 빅데이터를 활용한 연구가 활발히 이루어질 것으로 예상됨 (Athey, 2017)**

[참고1] 예측 모델링

모델링은 모델 설계자가 모델을 통해 답하고자 하는 질문과 연구 목적을 바탕으로 이루어짐

① **이론적 모델링**은 (1) 인과적 질문에 대한 수학적 표현 (2) 모델에 깔려 있는 가정을 상호 소통할 수 있도록 하는 정확한 수학적 언어를 제공 (3) 답하고자 하는 질문을 풀 수 있는 체계적인 틀을 제공(Pearl, 2009)해야 한다는 조건을 충족해야 함

: 경제변수들 간 **인과관계**에 대한 질문의 예로, "미국 연준의 기준금리 상승이 신흥국 자본유출을 야기하는가?" "기업세 인하는 기업의 투자를 촉진하는가?" 등을 들 수 있음

② **축약형 모델링**은 변수들 간의 상관관계에 기반하여 경제변수를 설명•예측

: **상관관계**란 두 변수의 선형 동행성(linear co-movement)에 기반한 유추로서, 예를 들어 올해 아프리카 지역의 강수량 추세($\mathbf{x_t}$)와 서울 지역 고등학생들의 평균 수학 점수 추세($\mathbf{y_t}$) 간 음적 상관관계 결과를 바탕으로 $\mathbf{x_t}$가 $\mathbf{y_t}$를 야기한다는 인과적 해석을 내리는 것은 명백한 오류

: 상관관계에만 기반해서는 인과관계를 도출해낼 수 없으며 인과관계는 인과적 가정에 의해 정의됨. 또한 인과관계는 상관관계와 달리 데이터를 통한 입증이 불가

[참고2] 예측력 평가

모델에 대한 평가는 모델 자체의 이론적 적합성뿐 아니라 모델의 예측력에 대한 상대적 비교 형태로 이루어지는데, 예측 모델링의 경우 경제변수 예측력을 제고하는 방법론을 선택

- 예측력에 대한 논의는 **실증 분석 결과에 기반한 결과론적 평가**로 이루어지는데, 그간의 실증 연구에 따르면 모든 경우의 수 -경제변수, 샘플 기간, 예측 기간(단기 • 장기)- 에 걸쳐 절대적으로 우세한 모델은 없다는 결론

- 즉, 문제 환경에 따라 예측력 결과가 다르게 나타나므로, 샘플기간 • 변수 • 특정 국가 혹은 산업 • 예측 기간에 따라 모델 간 비교 우위가 변동

- 하나의 최적 모델을 찾기보다는 각 모델에 상응하는 가중치를 부여하여 예측값들을 결합하는 방법(forecasting combination)도 많이 쓰임

2. 머신러닝 - 빅데이터 시대의 축약형 접근법

□ **머신러닝은 설계된 알고리즘을 따라 빅데이터에 담긴 정보를 학습하여 예측력 제고에 유용한 관심변수 정보를 추출하는 귀납적 러닝 메커니즘**

○ **머신러닝은 변수 간 관계 등과 같은 모델러의 사전적 지식(prior)을 요구하지 않으며, 주어진 알고리즘 틀을 이용해 데이터라는 Blackbox에 담긴 정보를 학습**

○ **머신러닝을 활용한 빅데이터 학습은 경제학뿐만 아니라, 금융·컴퓨터·공학·화학 등 여러 학문·산업 분야에 활발히 활용되고 있음**

- 특히, 연간별·분기별 데이터가 아닌 고빈도 데이터(high frequency: 매시,· 매분,· 매초)를 다루는 금융 공학과 기상학 분야의 경우 빅데이터를 활용한 머신러닝 기법의 중요성이 더욱 부각됨

- 각계 세계 중앙은행(NY Fed와 ECB 등)과 주요 테크기업(Google, Facebook, Amazon, Uber, Yelp, Airbnb 등)은 빅데이터를 예측 및 주요 의사결정에 적극 활용하고 있음

○ **머신러닝 알고리즘은 빅데이터에 담긴 정보를 학습하는 방법과 변수들에 가중치를 부과하는 방법에 따라 구별됨**

- 머신러닝의 장점은 다양한 구조를 가진 데이터를 다룰 수 있게 해주며, 기존 축약형 모델링의 문제점으로 지적된 변수의 자의적 선택 문제와 과적합(overfitting) 문제를 최소화

- 머신러닝은 빅데이터를 학습·요약하는 데 초점이 있는지(clustering, classification, dimension reduction) 또는 특정 종속변수의 예측성 제고에 도움이 되는 설명변수들을 추출하는 것인지에 따라 크게 ① **자율학습**(Unsupervised Learning)과 ② **지도학습**(Supervised Learning)으로 분류

□ **머신러닝은 데이터를 훈련시키는 효과적인 툴이지만, 연역적 사고 역시 모델링에 중요한 통찰력을 제공**

○ **모델점검·변수선택·결과해석 시 연역적 정보는 가이드라인으로서 역할**

- 머신러닝은 데이터가 담고 있는 이야기에 온전히 초점을 맞추기 때문에("a powerful tool to hear what data have to say"), 질적으로 유용한 정보를 사용하는 것이 무엇보다 중요하며 모델러의 직관도 적극 활용할 필요

2. 머신러닝 기법을 활용한 변수 예측

□ 실시간 데이터가 기록되는 빅데이터 시대에 머신러닝 기법을 활용한 예측의 산업체 활용방안은 무궁무진함

○ **최근 머신러닝에 대한 산업공학·물리학·통계학을 넘나드는 학제적 논의가 활발하지만, 실제 활용은 산업 현장의 주도로 이루어지고 있는 양상임**

- 2010년대에 들어 각광 받기 시작한 알고리즘들은 수십 년 전 학계에서 개발되었지만 활발히 활용되지 못하다가, 빅데이터와 하드웨어가 뒷받침됨에 따라 재조명되고 있음[3]

  · 주요 테크·서비스업체는 이용자들의 패턴을 학습, 은행·제조업은 상품·원료가격 및 주요 경제예측 등에 빅데이터를 적극 활용함에 따라 Data Scientist들을 적극 채용하고 있는 추세임

○ **머신러닝 기법을 활용하여 주요 의사결정에 빅데이터 정보를 스마트하게 활용 가능**

- 설명변수의 개수가 많은 예측 모델링의 경우, 지도(Supervised)학습 기법을 활용하여 종속변수와 상관성이 높은 변수 위주로 추출

  · 세계 주요 중앙은행들은 거시·금융변수 예측 시 관련 데이터를 모두 집산한 뒤 지도학습을 통해 좀 더 유용한 변수들을 추출하여 설명변수로 사용하는 방법을 채택

  · 일 예로 **철강 및 원료 수요 예측** 시 철강 및 원료가격, 국내외 거시/금융변수, 전후방 관련 산업체의 산업지수 및 영업이익 추이, 재고물량, 관세율 등을 관련 설명변수로 생각할 수 있는데, 지도학습을 통해 좀 더 유의미한 변수를 추출하거나 중요도에 따라 가중치를 부여 가능

- 뚜렷한 문제·예측변수가 주어지지 않더라도 자율(Unsupervised)학습 기법을 이용하여 데이터를 분류(classification) 및 군집(clustering)하여 데이터 패턴을 파악하고 빅데이터에 담긴 정보를 학습

  · 구글 서치 데이터를 기반으로 한 소비자의 구매 패턴·기호 분석과 같이 빅데이터를 활용한 상용화 방안은 생활 속에 밀접하게 침투되어 있음

  · 일 예로 매 순간 쌓이는 만 단위의 데이터를 통해 직원들의 behavioral

[3] 예를 들어, 통계학과 컴퓨터공학에서 시초한 Boosting과 Bagging은 1990년대에, 신경학자들이 개발한 Neural network는 1940년대에 처음 제안됨

pattern을 파악[4] 하여 **Smart한 기업경영**에 적용 가능. 빅데이터에 담긴 정보를 적극 활용하여 직원들의 복지 향상을 위한 정책과 더욱 효과적인 인사 인센티브 정책 고안

□ 빅데이터 방법론은 기존 데이터에 담긴 정보를 학습하는 데 초점을 맞추므로 데이터 자체가 편향된(biased) 경우 편향된 예측결과로 이어짐

○ **Amazon은 기존 직원들의 이력서에 담긴 정보를 머신러닝 기법으로 학습하여 인사채용에 활용하였지만 성차별적인 결과를 도출한다는 점을 발견하고, AI 채용 절차를 폐지** (Reuters, 2018.10.10일자)

- 테크 산업은 남성 엔지니어의 비중이 큰 분야인데, 기존 인력 풀의 정보 학습에 기반한 머신러닝 알고리즘이 이력서에 쓰인 여성 관련 단어에 페널티를 부과하였고, 이는 기존의 남성지배적 구조를 고착화하는 결과를 초래

- 맥락에 대한 이해가 뒷받침되지 않은 기계적 학습의 폐해를 단적으로 드러내는 사례로서 모델러의 조정적 역할이 더욱 강조됨

□ 빅데이터 시대에 모델러에게 요구되는 자질은 맥락에 맞는 데이터 활용 능력

○ **알고리즘은 정형화되고 슈퍼컴퓨터를 활용한 실시간 머신러닝도 가능해졌지만, 핵심은 문제 상황에 어떻게 적용하고 해석할 것인가임**

- 특히, 머신러닝 기법의 단점으로 지적되는 결과의 해석력과 다른 상황에의 적용 가능성 문제를 보완하기 위해서는 모델러의 역할이 더욱 강조됨

- 과거 정보 학습에 기반한 미래 예측이라는 프레임에 제약을 받기보다는, 변화하는 상황에 맞춰 유의미한 정보들을 유연하게 통합할 필요가 있음

  · 예를 들어, 2008년 금융위기 이후 금융규제망이 엄격해지고 새로운 양상의 그림자 금융이 생겨나게 되었는데, 이에 따라 금융 불확실성을 모델링하기 위해서는 신종 금융상품의 정보를 수치화하는 등 새로운 금융시장 생태에 맞는 모델링이 요구됨

  · 철강•원자재가격들은 날씨•기술•정부정책 등에 의한 공급 측 요인 변동 및 미중 지정학적 대립관계에서 비롯되는 교역환경 변화에 영향을 많이 받는 편. 이러한 변수들을 어떻게 더 체계적으로 수량화할 것인지는 중요한 문제이며 스마트한 도구 변수(instrument variable)를 선택해야 함

[4] 프라이버시 침해 문제를 예방하기 위해, 데이터를 익명화하는(de-identify) 선작업이 요구됨

3. 시사점

□ **모델의 예측력을 현 시점에서 평가하는 것은 현재 확보 가능한 데이터를 기반으로 할 수밖에 없다는 근본적인 한계가 있음**

○ **미래 추정값을 산출할 수는 있으나 실제값과의 비교는 불가능하므로, 경제예측력 평가에 대한 학술적 논의는 현재까지 주어진 데이터를 활용해 이루어짐**

- 현재 활용 가능한 샘플을 모델 fitting을 위한 in-sample과 예측력 평가를 위한 out-of-sample로 나누어 모델의 예측력을 평가

  · 예를 들어, 2008-2014년까지의 정보를 이용해 모델 모수값(parameter)을 추정하고(in-sample), 2015-2018년 기간(out-of-sample)에 해당하는 예측값을 산출한 뒤 실제 데이터와의 차이 값(forecast error)을 도출하여 모델의 예측력을 비교("pseudo-out-of-sample" forecasting evaluation)

- out-of-sample 기간 동안 가장 작은 예측오차 값을 산출해내는 모델을 예측력이 가장 우수한 모델로 선택

○ **즉, 실제 예측력 평가는 사후적(ex-post)으로 이루어질 수밖에 없고, 현재까지 확보 가능한 데이터로 순위를 매긴 예측력 평가가 미래에도 유효할지는 또 다른 문제**

- 경제예측은 변수들의 정상성(stationarity)을 가정하고 이루어지는데, 이는 큰 경제적 변동이 없는 이상 현재에 바탕한 결과가 미래에도 유의미할 것이라는 가정으로 모든 예측 모델링의 근본적인 한계이자 제약

○ **변수의 미래값을 예측하는 모델에는 현재 예상 및 관찰할 수 없는 갑작스러운 변수의 이상현상 및 경제충격은 반영되어 있지 않음**

- 현재 시점에 관찰 가능한 경제변수의 동태는 현존하는 의사결정 프레임에 의한 것이고, 현 시점 최적의 알고리즘은 지금까지 확보 가능한 데이터에 기반

- 모델 설계자로서 할 수 있는 최선은 변화된 상황에 맞게 알고리즘을 조정하고 가장 최신의 데이터를 활용하여 정보를 업데이트함으로써 끊임없이 조정

□ **적시성이 강조되는 산업현장의 경우 시시각각 변하는 현장의 정보와 모델 설계자의 통찰력·직관을 반영한 정성적 평가도 큰 몫을 담당**

○ **"economic forecasting is an art, not a science"라는 비판은, 경제예측의**

근본적인 한계를 지적하며 과학적 접근법에 결합되는 모델 설계자의 직관적 조정의 중요성을 강조

- 특히 경제학은 사회과학인 만큼, 기계화된 알고리즘을 통한 데이터의 효율적 학습도 중요하지만 정치경제적 상황에 대한 이해 및 산업 전문지식을 활용하여 모델을 맥락화하는 방안을 고안할 필요

- 데이터나 알고리즘에 편향(bias)이 있는 경우, 모델 설계자가 장기 추세를 참고하고 맥락에 맞는 해석을 통해 주어진 데이터와 모델이 설명하지 못하는 부분을 조정

  · 모델러의 사전 지식과 질적인 정보를 prior로 활용하는 Bayesian Factor Model과 Bayesian VAR도 하나의 대안

○ **예측력을 강화하기 위해서는 시시각각 변하는 정치경제적 상황과 각 산업의 특수성을 고려한 미세 조정(fine-tuning)이 필요**

- 빅데이터를 활용할 수 있는 프레임이 주는 장점을 활용하되, 해당 알고리즘의 장단점을 잘 인지하여 실시간 데이터와 모델을 업데이트할 필요

[참고자료]

[보고서/논문]

Athey, S., "The impact of machine learning on economics", *Economics of Artificial Intelligence,* University of Chicago Press, 2017

Refet S. Gü rkaynak, Burç in Kısacıkoğlu, Barbara Rossi, "Do DSGE models forecast more accurately out-of-sample than VAR models?", *VAR Models in Macroeconomics - New Developments and Applications: Essays in Honor of Christopher A. Sims (Advances in Econometrics, Volume 32),* Emerald Group Publishing Limited, pp.27-79, 2013

Pearl, J., "Causal inference in statistics: an overview", *Statistics Surveys,* v.3, pp.96-146, 2009

Varian, H. R., "Big data: new tricks for econometrics", *Journal of Economic Perspectives,* v.28 n.2 pp.3-28, 04/2014

Wolpert, David H. "The lack of a priori distinctions between learning algorithms", *Neural Computation,* v.8 n.7, pp.1341-1390, 1996

[서적]

Goodfellow, Ian, et al. *Deep learning*. v.1, Cambridge: MIT press, 2016

[언론]

Reuters, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women", October 10, 2018

Risk.net, "BlackRock shelves unexplainable AI liquidity models", November 12, 2018