Journal of KIIT. Vol. 22, No. 2, pp. 21-33, Feb. 28, 2024. pISSN 1598-8619, eISSN 2093-7571 21
http://dx.doi.org/10.14801/jkiit.2024.22.2.21

재귀적 베이지안 앙상블 모델링 기법을 이용한 사회경제지표의 다변량 예측

변준현*[1], 민수홍*[2], 강지훈**

Multivariate Prediction of Socio-Economic Indicators using Recursive Bayesian Ensemble Model

Junhyun Byun*[1], Suhong Min*[2], and Jihoon Kang**

본 논문은 2022년 과학기술정보통신부의 지원[2020R1F1A1074947 13]과 2023년 과학기술정보통신부의 재원으로 정보통신산업진흥원의 지원[S0317-21-1002]에 의한 연구임

요 약

코로나 펜데믹 이후로 국내외 사회경제의 불확실성이 증가하고 있으며, 이에 효과적인 대응을 위한 사회경제지표 예측은 필수적인 상황이다. 기존에는 사회경제지표 예측 방법으로 방정식, 시계열, 머신러닝 모델 등이 주로 사용되었으나 사회경제지표의 특징과 예측모델의 한계점 등으로 인해 저조한 예측성능을 보이고 있다. 이에, 본 연구에서는 재귀적 베이지안 앙상블 모델을 토대로 시간 흐름에 따른 불확실성이 큰 사회경제 지표를 안정적으로 예측하는 방법론을 제안한다. 제안모델은 예측원리가 다른 시계열, 머신러닝 등의 다양한 예측모델을 확률적으로 조합하고, 베이지안 업데이트 기법을 통해 모델을 순차적으로 업데이트한다. 본 연구는 사회경제지표 예측을 통해 제안모델의 예측 성능과 안정성의 우수함을 입증하였다.

Abstract

After the COVID-19 pandemic, there has been an increase in uncertainty in both domestic and international socio-economic conditions, highlighting the necessity of accurate predictions for socio-economic indicators(SEIs). Historically, equations, timeseries, and machine learning models have been widely used for forecasting SEIs, but, their predictive accuracy has been constrained by the characteristics of SEIs and limitations in prediction models. Thus we proposed the Recursive Bayesian Ensemble Model(RBEM) to reliably predict SEIs that exhibit high uncertainty over time. The proposed model probabilistically combines various prediction models with different forecasting principles, such as timeseries models and machine learning, and sequentially adopts recursive Bayesian update techniques. The predictive performance and stability of the proposed model have been demonstrated through the prediction of SEIs.

Keywords
socio-economic indicator, ensemble, fitness weight, recursive bayesian update

* 한국공학대학교 스마트팩토리융합학과 석사과정
- ORCID[1]: https://orcid.org/0000-0002-0212-2741
- ORCID[2]: https://orcid.org/0009-0007-0573-6630
** 한국공학대학교 경영학부 조교수(교신저자)
- ORCID: https://orcid.org/0000-0003-3784-5958

· Received: Dec. 04, 2023, Revised: Jan. 15, 2024, Accepted: Jan. 18, 2024
· Corresponding Author: Jihoon Kang
Dept. of Business Administration, Tech University of Korea
237, Sangidaehak-ro, Siheung-si, Gyeonggi-do, Republic of Korea
Tel.: +82-31-8041-0687, Email: jhkang82@tukorea.ac.kr

Ⅰ. 서 론

사회경제지표(Socio-economic indicators)는 사회, 경제에 관한 계량 통계치를 의미하며, 각 지표들은 사회, 경제 현상에 대한 설명과 복합적인 인과관계 이해를 위해 사용된다[1]. 대표적으로 계량경제학(Econometrics)은 사회경제지표들을 활용하여 경제 현상의 해석, 미래에 대한 예측, 경제 정책의 분석을 수행하는 학문으로 정부, 기업의 의사결정 과정에 관여해왔다[2]. 최근 국내외 경제의 불확실성이 증가하고 있으며, 이에 효과적인 대응을 위해 미래 경제상황에 대한 더욱 정밀한 예측이 필요한 실정이다. 이에 따라 객관적인 미래 지표 예측을 위해 사회경제 데이터를 분석 및 예측하는 계량경제 연구들이 지속되고 있다[3]-[5].

사회경제지표 예측 방법은 크게 방정식 기반(수리모델), 시계열 모델 기반, 머신러닝 기반으로 나뉜다. 수리 모델 기반의 방법론은 가장 전통적인 계량분석 방법으로, 연구자가 사회경제에 대한 사전지식을 기반으로 경제 변수들을 활용하여 경제 모형을 설계한 후, 각 변수의 핵심 영향도인 계수(Coefficient)들을 추정하는 방법이다. 모델이 주로 수식으로 구성됨에 따라, 경제 이론적 지식을 토대로 모형 및 모형의 결과에 대한 해석이 가능하다는 장점이 있지만, 모델의 설계 과정에서 연구자의 주관과 사전지식에 주로 의존해야 하는 한계점이 존재한다[6].

이러한 방정식 기반의 모델의 한계를 극복하기 위해, 일정한 시간 간격에 따라 관측된 자료인 시계열 자료를 분석하여 미래를 예측하는 시계열 분석 모형이 제안되었다[7][8]. 시계열 분석 모형은 시간(t)에 따른 종속변수의 추세(Trend) 및 다양한 주기적 속성(Seasonality) 등을 이용하여 미래의 종속변수를 예측하는 방법으로써, 대표적인 모델로는 ARIMA(Auto-Regressive Integrated Moving Average) 모형이 있다[9]. 시계열 모델은 모델 구축에 필요한 분야별 전문 지식과 이를 수학적으로 모델에 반영하려는 노력이 방정식 모델에 비해 상대적으로 적은 장점이 있다. 하지만, 종속 시계열 변수 자체가 정상 확률 과정을 만족해야 하고, 만족하지 않는 경우 모델의 계수 추정이 불확실해지며, 또한 자기상관이 변동되어 의미 있는 모델링이 어려워진다. 이러한 시계열 모델의 특성은 대부분의 변수가 비정상 시계열 데이터인 사회경제지표의 미래 예측에서 불리하게 작용한다[10].

추가적으로 다음과 같은 사회경제지표의 특성으로 인해 예측은 더욱 어려워진다. 먼저, 민간 소비 증가에 따른 기업의 고용 증가와 같이 사회경제지표들 간의 선후 상관관계가 존재하며, 이러한 관계를 식별하고 이해하는 것은 예측 문제에서 중요한 역할을 수행한다[11]. 이와 같은 관계를 적절하게 반영하지 않은 모델은 미래 예측성능이 낮아지게 된다. 또한, 사회경제지표의 원 현상인 사회경제상황이 지속적으로 변화함에 따라, 사회경제지표는 시기별로 새로운 패턴이 등장한다는 점이다. 이로 인해 특정시점에서 학습된 모델은 시간이 지남에 따라 달라지는 데이터 패턴을 반영하지 못하게 되어, 모델의 예측력 저하 및 불확실성 증가를 야기한다. 더욱이 최근 사회경제지표를 예측하는데 있어 고려해야하는 변수들이 증가하고, 비선형성이 증가함에 따라 기존의 방정식 기반, 시계열 모델 기반 방법들만으로 미래 상황을 정확하게 가늠하는 것은 매우 어려운 일이 되었다[12]-[14].

이러한 다변수 및 비선형성 문제들을 해결하기 위해 머신러닝을 활용하여 사회경제지표를 예측하는 연구들이 다양한 형태로 활발히 연구되고 있다[15][16]. 이주미 외[5]는 부동산 지수인 아파트 매매지수, 지가지수, 전세가격지수, 부동산 심리지수 데이터를 활용하여 LSTM 모델과 머신러닝 모델인 랜덤 포레스트, XGBoost 모델을 통해 각 지수별 단변량 예측을 수행하였다. 실험 결과 부동산 지수의 예측은 데이터의 주기 특성과 형상에 따라 머신러닝 모델별 예측 정확도가 상이함을 보였다. 이렇듯, 머신러닝 기반의 연구들은 다양한 사회경제 분야에서 기존 계량경제학 연구 방법들 대비 우수한 예측성능을 보였다. 하지만 머신러닝 기반의 계량경제 모델링 연구는 종속변수의 특징 및 예측 가능 기간마다 모델별 예측 정확도가 변동되며, 이에 따라 머신러닝 모델들의 예측성능 분산이 커지는 한계를 갖고 있다. 이러한 머신러닝의 한계점들을 극복하기 위해 다양한 머신러닝 모델의 예측 결과를 조합하는 방법론에 대한 연구들이 진행되었다.

대표적으로 J. W. Kim[17]은 산업별 부가가치율을 다양한 머신러닝 기법들을 활용하여 예측하였다. 예측의 안정성을 위해 각 머신러닝 기법들의 예측값을 조합하여 산업별 부가가치율 예측을 수행하였다. 그 결과, 단일 머신러닝 대비 평균적으로 뛰어난 예측성능을 보였으나, 데이터 내부적으로 불규칙한 추세를 포함하는 경우에 예측결과의 분산이 커진다는 한계를 갖고 있다.

본 연구에서는 기존 계량경제 모델과 시계열 모델의 태생적인 서로의 한계를 극복하고 더욱 효율적인 사회경제지표 예측을 목표로 기존 계량경제 모델과 시계열 모델의 태생적인 서로의 한계를 극복하고, 머신러닝 모델에서 발생 가능한 과적합(Overfitting) 문제를 병렬적으로 해결하기 위해 복수의 예측모델을 적합도 가중치에 따라 조합해서 예측하고, 모델을 재귀적으로 업데이트하는 방식인 재귀적 베이지안 혼합형 모델을 제안한다. 본 논문의 구성은 다음과 같다. 제 2장은 연구동기와 수집한 사회경제지표에 대해 설명한다. 제 3장은 제안 방법론을 설명한다. 제 4장은 제안방법론과 제안 방법론의 구성 모델들과의 사회경제지표 예측 성능 비교 및 검증 결과를 제시한다. 제 5장에서는 본 연구의 결론에 대한 설명과 향후 연구방향을 제시한다.

II. 데이터 수집 및 연구 동기

2.1 데이터 수집

본 연구에서 수집한 지표는 표 1과 같다. 수집한 사회경제지표 데이터는 모두 공개된 데이터로, 주요 데이터 출처는 다음과 같다. 인구통계정보는 행정안전부(https://jumin.mois.go.kr), 가계대출과 금리 관련 데이터는 e-나라지표(https://www.index.go.kr), 부동산 관련 데이터는 한국부동산원 부동산통계정보시스템(https://www.reb.or.kr/r-one), 일반적인 거시경제지표들은 한국은행경제통계시스템(https://ecos.bok.or.kr)에서 데이터를 확보하였다. 데이터 수집 기간은 2012년 1월 ~ 2020년 12월까지, 변수는 총 37개로 월별 분기별 기준으로 총 114개월 분량의 데이터를 수집하였다.

사회경제지표는 지표에 따라 월별 혹은 분기별로 수집되며, 데이터가 수집되는 주기가 상이한 특징을 갖는다. 데이터 기반의 신뢰성 있는 예측 모델링을 수행하기 위해서는 학습(Training)에 많은 양의 데이터가 필요하다[18]. 하지만 분기별로 수집되는 사회경제지표를 예측 모델링에 활용할 시 데이터가 근본적으로 부족하다는 문제점을 갖고 있다.

따라서 효과적인 모델링을 위해, 본 연구에서는 분기별로 수집되는 지표의 경우, 월별 데이터를 구성할 때 해당 분기별 데이터와 동일한 값으로 보간하여 분기별 지표를 월별지표로 변환하였다.

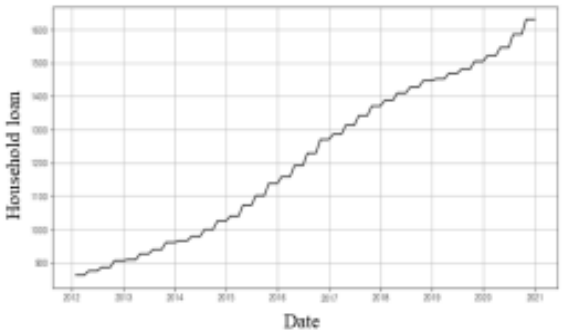
표 1. 수집한 사회경제 지표
Table 1. Collected socio-economic indicators

Variable name	Scope	Frequency	Unit
Total population	Nationwide	Quarterly	Persons
GDP	Nationwide	Quarterly	1 billion KRW
GNP	Nationwide	Quarterly	1 billion KRW
GDI	Nationwide	Quarterly	1 billion KRW
Household credit	Nationwide	Quarterly	1 trillion KRW
Household loan	Nationwide	Quarterly	1 trillion KRW
Sales credit	Nationwide	Quarterly	1 trillion KRW
Economic growth rate	Nationwide	Quarterly	%
Private consumption growth rate	Nationwide	Quarterly	%
Consumer price index	Nationwide	Monthly	Index
Consumer confidence index	Nationwide	Monthly	Index
Total savings rate	Nationwide	Quarterly	%
Gross domestic investment rate	Nationwide	Quarterly	%
Unemployment rate	Nationwide	Monthly	%
M2 (Seasonally adjusted, Average balance)	Nationwide	Monthly	1 billion KRW
M2 (Average balance) growth rate	Nationwide	Monthly	1 billion KRW
Government bonds (3 year treasury)	Nationwide	Monthly	%

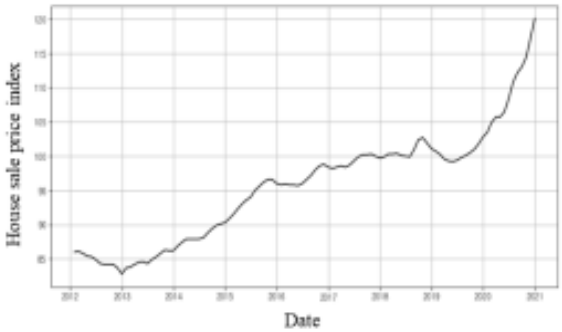
Government bonds (5 year treasury)	Nationwide	Monthly	%
Government bonds (10 year treasury)	Nationwide	Monthly	%
Corporate bonds	Nationwide	Monthly	%
CD 91	Nationwide	Monthly	%
Call rate	Nationwide	Monthly	%
Base rate	Nationwide	Monthly	%
Won-Dollar exchange rate	Nationwide	Monthly	KRW
Won-Yen exchange rate	Nationwide	Monthly	KRW
Won-Euro exchange rate	Nationwide	Monthly	KRW
KOSPI	Nationwide	Monthly	Index
KOSDAQ index	Nationwide	Monthly	Index
Actual transaction price index for apartment houses	Nationwide	Monthly	Index
Apartment transaction volume	Nationwide	Monthly	Units
Unsold housing units	Seoul	Monthly	Units
Housing construction license performance	Seoul	Monthly	Units
Housing completion performance	Seoul	Monthly	Units
Housing construction performance	Seoul	Monthly	Units
Apartment Sale price	Nationwide	Monthly	1,000 KRW
Jeonse price rate	Nationwide	Monthly	%
House sale price index	Nationwide	Monthly	Index

2.2 연구 동기

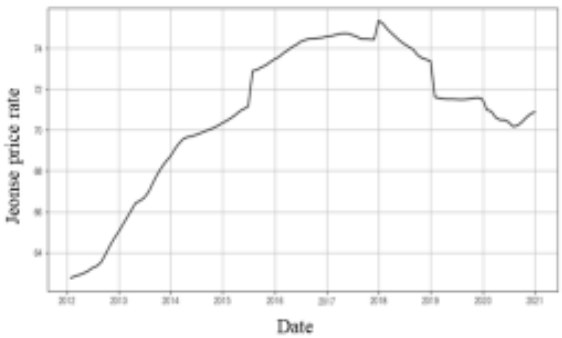
그림 1은 일부 사회경제지표를 시각화한 결과다. 수집 기간 내에서 그래프를 확인했을 때, 가계대출은 선형적인 우상향 패턴, 종합주택매매가격지수는 비선형적인 우상향 패턴, 전세가율은 상승과 하향 패턴이 공존하는 비선형 패턴, 실업률은 강한 계절성 패턴을 띤다. 이는 상이한 패턴을 띠는 지표를 예측에 활용할 시 각기 수학적으로 다른 표현이 필요할 것으로 보인다.



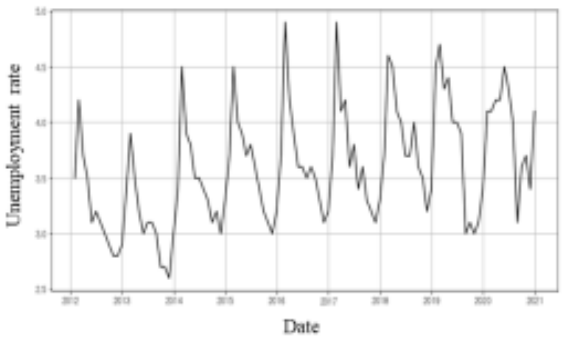
(a) 가계대출
(a) Household loan



(b) 종합주택매매가격지수
(b) House sale price index



(c) 전세가율
(c) Jeonse price rate



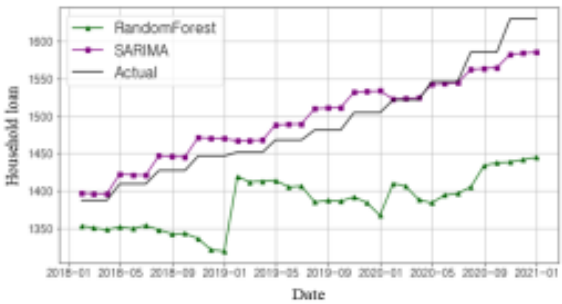
(d) 실업률
(d) Unemployment rate

그림 1. 사회경제지표 시각화
Fig. 1. Visualization of the socio-economic indicators

수집한 사회경제 지표들을 활용하여 머신러닝 모델을 통해 예측하였을 때, 모델별로 서로 다른 예측 성능을 나타내고 있다.

그림 2는 다변량 머신러닝 모델에서 대표적으로 활용되는 랜덤 포레스트(Random forest) 모델과 시계열 모델인 SARIMA(Seasonal ARIMA) 모델로 가계대출과 실업률 변수를 예측한 결과다. 그림 2에서 활용된 종속변수는 가계대출과 실업률을 사용하였고, 독립변수로는 종속변수를 제외한 총인구, GDP 등의 36개 변수가 사용되었다.

그림 2(a)에서 트렌드 성분이 명확한 가계대출의 경우 다변량 머신러닝 모델보다 단순 시계열 예측 모델이 더욱 우수한 예측을 수행함을 볼 수 있다. 반면에 그림 2(b)에서 시계열적으로 강한 계절성과 다른 변수들과 전형적 관계를 갖는 패턴인 실업률의 경우 단순 시계열 예측모델보다 다변량 머신러닝 모델이 장기예측 관점에서 예측력이 더욱 뛰어난 것을 볼 수 있다. 이처럼 특정 변수, 특정 모델에 따라 성능 차이가 존재하는 것을 볼 수 있으며, 이는 예측원리가 다양한 예측모델이 결합된 형태의 모델링 기법 도입 필요성을 나타낸다.



(a) 가계대출
(a) Household loan

(b) 실업률
(b) Unemployment rate

그림 2. 사회경제 지표에 대한 다양한 모델의 예측 결과
Fig. 2. Predictive results of various models for socio-economic Indicators

III. 제안 방법론: 재귀적 베이지안 앙상블 모델

본 논문은 단일 예측의 한계점과 사회경제지표의 패턴 불규칙성으로 인한 모델링 불확실성을 보완하고자 재귀적 베이지안 앙상블 모델 (RBEM, Recursive Bayesian Ensemble Model)을 제안한다. RBEM 모델은 예측 원리가 다른 다양한 예측모델과 예측 모델별 성능의 우수성을 나타내는 지표인 적합도 가중치(Fitness weights)를 조합하여 모델을 재귀적으로 예측 및 업데이트를 반복 수행한다.

제안 모델의 핵심 기법인 앙상블 예측과 베이지안 업데이트 방법은 특정 시기별 변화하는 패턴과 불규칙한 비선형 패턴을 보이는 사회경제지표 예측에 적합한 특징을 갖는다. 앙상블 예측의 이점은 각 예측모델이 서로 다른 모델링 체계를 가지고 있기에, 이를 혼합해서 사용할 경우, 다양한 패턴 예측에 있어서 상호 보완이 가능하다는 이점을 갖는다. 베이지안 업데이트 기법은 모델링 불확실성을 순차적으로(Sequentially) 줄여주는 기법으로, 예측모델이 과거의 데이터, 즉, 사전정보와 현재의 데이터인 관측 정보를 활용하여 현재 상태를 더욱 정확하게 추정할 수 있도록 도와준다[19].

그림 3은 RBEM 모델의 전반적인 예측 프로세스를 나타내며, RBEM 모델을 구축하기 위한 세부 과정은 다음과 같다: 3.1 초기 예측모델 설정, 3.2 적합도 가중치 계산, 3.3 앙상블 예측, 3.4 재귀적 베이지안 업데이트 적용.

3.1 초기 예측모델 설정

RBEM 모델은 복수의 예측모델을 조합(앙상블)하여 사용하는 예측방법으로, 우선 초기(Baseline) 예측모델을 설정해야 한다. 본 연구에서는 사회경제지표가 선형적, 비선형적, 주기적 패턴을 보임에 따라 머신러닝에서 선형패턴을 예측하는데 활용되는 능형 회귀모델(Ridge regression), 비선형 패턴을 예측할 때 쓰이는 랜덤 포레스트 모델, 시계열 예측에 활용되는 SARIMA 모델을 초기 설정하였다.

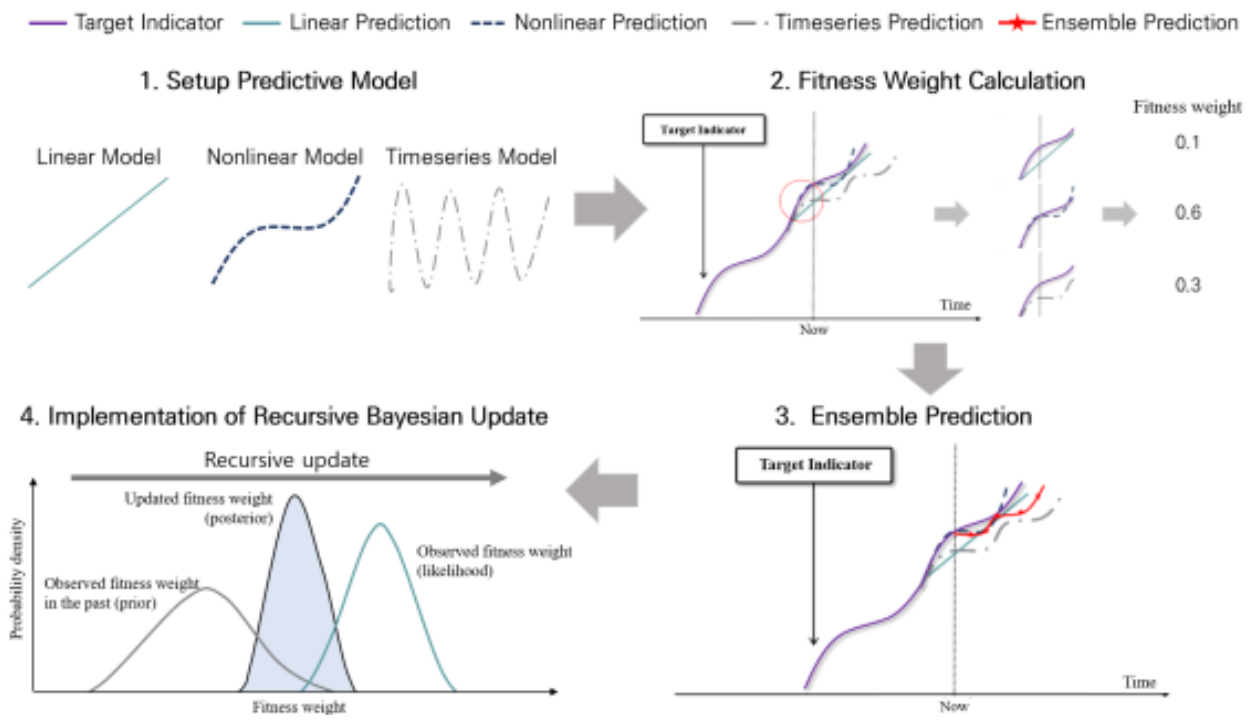


그림 3. 재귀적 베이지안 앙상블 모델 예측 프로세스
Fig. 3. Overall process of recursive Bayesian ensemble model

본 연구에서는 위 모델을 조합해서 활용하였으나, 향후 다양한 문제에서 특정 문제에 부합하는 모델을 사용한다면 안정적인 예측 결과를 얻을 것으로 사료된다.

3.2 적합도 가중치 계산

정확성과 안정성이 높은 앙상블 모델을 구축하기 위해, 각 예측모델의 가중치를 효과적으로 부여하는 것이 모델 조합에서는 핵심적인 부분이다. RBEM모델의 적합도 가중치는 각 예측모델의 예측 성능을 기반으로 계산된다. 우선 RBEM모델은 학습데이터를 활용하여 사전에 설정한 예측모델을 구축 후 평가데이터를 통해 예측 결과에 대한 잔차(Residual)를 계산한다. 그 후, 각 모델의 잔차는 적합도 가중치로 사용하기 위해 잔차의 역수로 변환하고, 적합도 가중치를 0부터 1사이로 정규화하여 확률적으로 표현한다. 적합도 가중치의 산출은 식 (1)을 통해 표현된다.

$$w_m = 1 / \frac{r_m}{\sum_m^M r_m}, \quad 0 \le w_m \le 1 \tag{1}$$

r_m $m = 1, 2, \ldots, M$은 앙상블 모델에서 사용되는 예측모델별 잔차를 뜻하며, w_m은 각 모델의 적합도 가중치를 나타낸다.

3.3 앙상블 예측

RBEM 모델은 각 모델의 예측값과 예측오차 기반의 적합도 가중치의 가중합을 통해 최종 예측값을 산출한다. RBEM 모델의 최종 예측값은 식 (2)와 같다.

$$\hat{y} = \sum_m^M \hat{w}_m \hat{y}_m \tag{2}$$

$\hat{w}_m$은 사전정보가 반영된 예측모델별 적합도 가중치를 나타내며, $\hat{w}_m$에 대한 설명은 재귀적 업데이트에서 자세하게 다룬다. $\hat{y}_m$은 예측모델별 예측값을 의미한다.

3.4 재귀적 베이지안 업데이트

RBEM 모델에서 적합도 가중치는 베이지안 업데이트(Bayesian update) 기법을 통해 주기적으로 갱신된다. 본 연구에서 정의한 베이지안 업데이트는 수식 (3)과 같다.

$$p(\hat{w}_{m,t}|y_{m,t},\hat{y}_{m,t}) \propto \prod_{m}^{M} p(w_{m,t}) | L(w_{m,t}|y_{m,t},\hat{y}_{m,t}) \tag{3}$$

$p(w_{m,t})$는 사전확률(Prior)로 초기에는 성능이 우수한 모델을 알 수 없기 때문에 초기값을 $\frac{1}{m}$로 동일(Uniform)하게 설정한다. $L(w_{m,t}|y_{m,t},\hat{y}_{m,t})$는 우도(Likelihood)로 현재시점 평가데이터의 잔차를 활용하여 계산한다. $p(\hat{w}_{m,t}|y_{m,t})$은 사후확률(Posterior)로, 사전확률과 우도의 조합을 통해 보정된 적합도 가중치를 나타낸다. RBEM 모델의 적합도 가중치는 식 (3)의 과정을 통해 주기적으로 보정되고, t 시점의 사전확률은 시점이 지날때마다 이전 시점의 사후확률로 업데이트 된다.

Ⅳ. 사회경제지표 예측 실험

4.1 예측대상 지표 설정

사례연구에서는 가계대출, 부동산 지표, 실업률을 종속변수로 설정하였다. 가계대출은 가계부채의 빠른 증가는 금융안정 위험이 경제적 성장을 저해할 수 있다[20]. 부동산 지표는 국가, 기업, 가계가 보유한 자산중에서 가장 큰 비중을 차지하는 요소이며, 부동산의 큰 가격변동은 금융시장의 안정성 저하를 유발할 수 있다[21]. 또한 실업률은 노동시장에 따른 경제 변화에 국가정책이 탄력적으로 대응할 수 있도록 도움을 준다[22].

4.2 예측수행 방법

본 장에서는 위에서 종속변수로 설정한 가계대출, 종합주택매매가격지수, 전세가율, 실업률을 예측하였다. 예측모델에 활용되는 학습 기간은 2012년 1월 ~ 2017년 12월이며 2018년 1월 ~ 2020년 12월 데이터를 활용해 RBEM 모델을 평가하였다. 또한, 사회경제지표가 갖는 선후관계, 시간변동성 문제를 완화하고자 그레인저 인과관계 검정을 통한 사회경제지표 간 선후관계 고려, 무빙윈도우 기반 예측 모델링을 수행하였다.

그레인저 인과관계 검정 방법은 종속변수의 과거값만으로 종속변수를 예측하는 것보다 독립변수의 과거값을 같이 활용했을 때 예측성능이 더 우수하다면 인과방향이 존재한다고 판단하는 인과관계 검정 방법이다[23]. 본 연구에서는 사회경제의 불확실한 상황을 효과적으로 대비하기 위해 최소 3개월 이전의 관계를 고려하여 그레인저 인과관계 검정을 수행하였다. 표 2는 그레인저 인과관계 검정 방법론을 사회경제지표에 적용한 결과 예시를 나타낸다. 표 2에서 평균전세가율을 설명함에 있어 5시점 전의 소비자물가지수가 평균전세가율에 영향을 주었다고 해석할 수 있다.

표 3은 그레인저 인과관계 검정 결과값을 활용하여 선후 상관관계를 반영한 데이터셋이다. 전세가율과 같이 가계대출, 종합주택매매가격지수, 실업률 지표도 동일하게 그레인저 인과관계검정 방법론을 적용하여 표 3과 같이 데이터셋을 구축 후 예측모델링을 수행하였다.

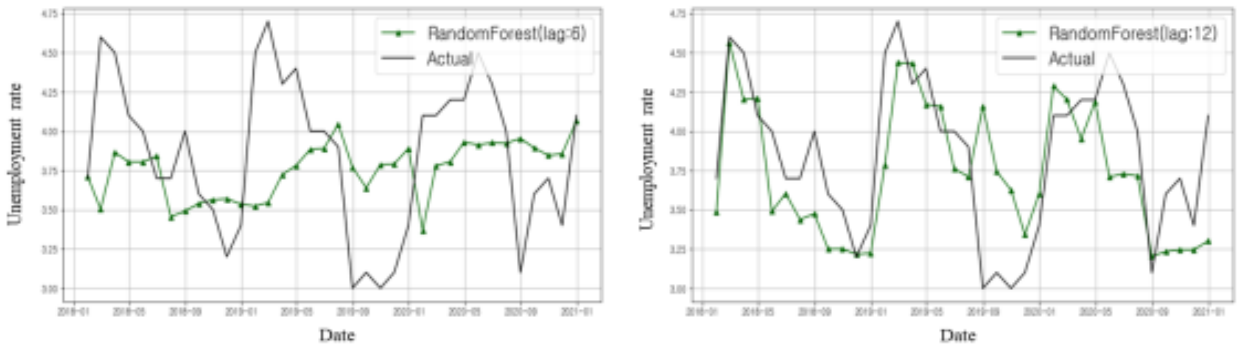
시계열 예측 모델링은 시차값에 따라 예측 결과가 상이한 특징을 가진다. 예를 들면, 그림 4는 랜덤 포레스트 모델을 활용하여 시차를 다르게 적용해서 실업률을 예측한 결과다. 12개월마다 주기성이 명확한 실업률의 경우 시차를 12로 줬을 때 준수한 성능을 내지만, 12개월마다 주기성 패턴이 있는 데이터에 시차를 6으로 줬을 때 예측이 어려움을 보인다. 이는 더욱 정확한 예측을 하기 위해서는 최적의 시차를 탐색하는 과정의 필요성을 나타낸다. 본 논문에서는 예측 모델링의 정확도를 높이기 위해 변수별로 최적의 시차를 휴리스틱하게 탐색하였다. 표 4는 변수별 성능이 우수했던 시차값을 보여준다. 사회경제지표 예측은 그레인저 인과관계 검정 방법론을 통해 선후 상관관계를 반영할 때 표 4에 나타나는 최적의 시차를 적용하여 예측을 수행하였다.

표 2. 그레인저 인과관계 검정 적용 결과 예시 (평균 전세가율에 대한 검정 결과)
Table 2. Example of applying the granger causality test (test results for the average rental deposit rate)

Min lag	P-value	Series 1		Series 2
4	0.002	Call rate		
4	0.018	Total population		
4	0.213	Government bonds (3 year treasury)		
⋮	⋮	⋮	→	Price rate
5	0.044	Consumer price index		
3	0.058	Corporate bonds		
3	0.105	Sales credit		
3	0.110	KOSDAQ index		

표 3. 그레인저 관계 검정 방법론을 통해 산출된 선후관계를 반영한 데이터셋 예시(종속변수: 평균 전세가율)
Table 3. Dataset example reflecting the causal relationship derived through the Granger causality test methodology (dependent variable: average rental deposit rate)

Date	KOSDAQ index (lag:3)	Consumer price index (lag:5)	…	Price rate
2012. 01	NA	NA	…	62.7637
2012. 02	NA	NA	…	62.8431
⋮	⋮	⋮	⋮	⋮
2020. 11	865.53	104.87	…	70.7664
2020. 12	832.64	104.86	…	70.8764



(a) 실업률(시차:6)
(a) Unemployment rate (lag:6)

(b) 실업률(시차:12)
(b) Unemployment rate (lag:12)

그림 4. 시차(lag)별 실업률 예측 결과 비교
Fig. 4. Comparison of unemployment rate prediction results by time lag

표 4. 사회경제지표별 최적의 시차
Table 4. Optimal time lag for each main socio-ecnomic indicator

Socio-economic indicator	Household loan	House sale price index	Price rate	Unemployment rate
Optimal forecastable period(months)	12	6	3	12

무빙윈도우 기법은 그림 5와 같이 묘사될 수 있다. 무빙윈도우는 윈도우 사이즈 내 사회경제지표를 학습 후, 평가하고 시점이 지날 때마다 윈도우를 이동하여 반복적으로 학습 및 평가하는 방법이다. 사회경제지표 예측에서 활용된 초기 학습구간은 2012년 1월 ~ 2017년 12월 데이터로 60개의 데이터를 활용하였으며, 평가 구간은 표 4의 각 모델별 최적 시차 값으로 활용하였고, 2020년 12월까지 윈도우를 이동시켜 반복적으로 예측을 수행하였다. 무빙윈도우 기법은 시간이 지남에 따라 오래된 데이터는 제거되고 최신의 데이터를 활용해 주기적으로 사회경제지표를 예측하여, 실시간 예측에 이점을 갖는다.

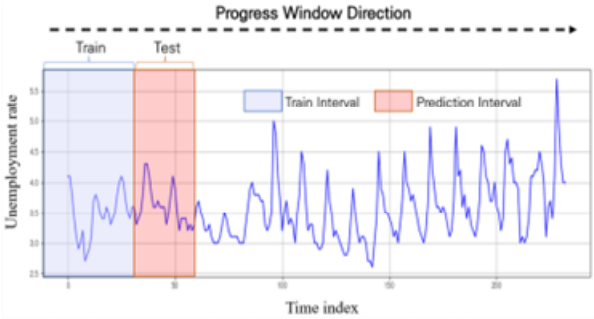


그림 5. 무빙윈도우 기법
Fig. 5. Moving window method

4.3 성능 지표

RBEM 모델의 예측성능을 측정하는 지표로는 MSE (Mean Squared Error)를 사용하였으며 식 (4)와 같다:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{4}$$

y_i는 실제 값을 나타내며, $\hat{y}_i$는 예측된 값을 나타낸다. MSE는 실제값과 예측값의 차이의 평균을 통해 성능을 정량적으로 측정한다.

4.4 실험 결과

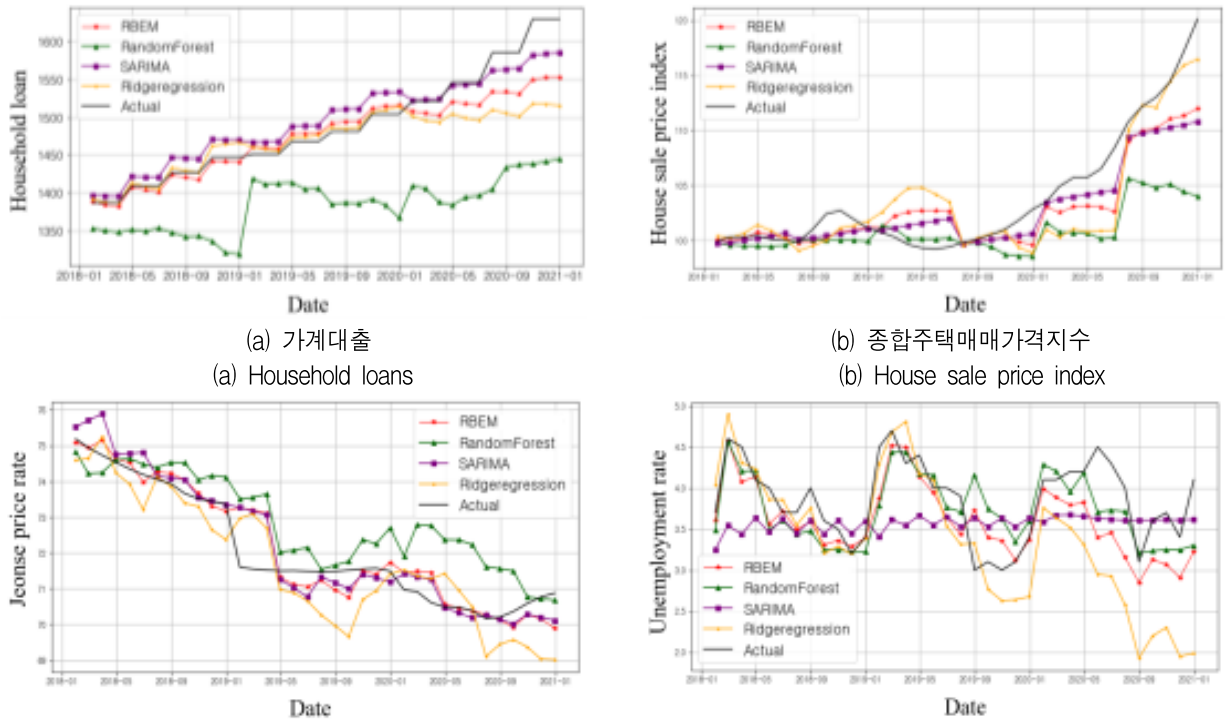
RBEM 모델의 성능평가는 RBEM의 구성모델인 랜덤 포레스트 모델, 능형 회귀모델, SARIMA 모델을 비교하여 성능을 평가하였다. RBEM 모델의 정확한 성능평가를 위해 정량적인 지표인 MSE 관점과 정성적인 그래프 관점에서 성능평가를 수행하였다. 그림 6은 그래프 관점에서 가계대출, 종합주택매매가격지수, 전세가율, 실업률에 대한 예측 결과를 보여준다. 그림 6을 보면, 사회경제지표별 예측 모델의 성능은 일관적이지 않다. 반면에 RBEM 모델은 대부분의 데이터 패턴에 근접한 모습을 보이며, 개별 모델 대비 비교적 예측 분산이 적은 것으로 보인다. 이것은 다양한 모델을 확률적으로 조합해서 사용하는 방법이 단일 모델을 사용하는 것보다 예측이 안정적이다는 것에 의의가 있다. 하지만 2020년은 예측이 어려움을 보이는데, 코로나 펜데믹으로 사회경제 상황의 급변으로 인한 데이터 간 상관관계의 변화로 발생하는 문제로 사료된다.

표 5는 가계대출, 종합주택매매가격지수, 전세가율, 실업률의 예측 결과에 대한 MSE를 나타낸다. 표 5를 보면, 가계대출과 종합주택매매가격지수는 SARIMA 모델의 성능이 1위로 가장 우수하며, 전세가율은 능형 회귀모델이 가장 우수하며, 실업률은 랜덤 포레스트 모델이 가장 우수한 것으로 나타났다.

RBEM 모델의 경우, 성능평가 비교를 위해 사용된 구성 모델의 평균치보다 모든 경우에서 더 뛰어난 성능을 보였다. 비록, 지표별 가장 에러가 작은 타 모델 대비 RBEM 모델이 큰 오차를 보였으나, 모든 지표에서 2번째로 높은 성능을 보였으므로 대부분의 경우에 안정적이고 최적에 가까운 결과를 나타낸다.

MSE 관점과 그래프 관점의 예측 결과를 종합적으로 고려해 보면, RBEM 모델은 기존의 단일 시계열 및 머신러닝 모델과 비교하여 상대적으로 안정적인 예측 결과를 보여주었다. 이러한 안정성은 주요 사회경제 지표를 예측하는 시뮬레이션 기법으로 RBEM 모델을 적절하게 활용할 수 있다는 가능성을 시사한다.

그러나 RBEM 모델은 데이터 기반의 예측 모델로, 코로나 펜데믹과 같이 예기치 못한 사건에서는 취약한 특징을 보인다. 따라서 정확한 예측을 위해서는 데이터 기반 모델뿐만 아니라 다양한 관점에서의 분석이 필요하다는 점을 나타낸다.



(a) 가계대출
(a) Household loans

(b) 종합주택매매가격지수
(b) House sale price index

(c) 전세가율
(c) Price rate

(d) 실업률
(d) Unemployment rate

그림 6. 주요 사회경제지표에 대한 예측 결과 비교 (Random forest, SARIMA, ridge regression, RBEM)
Fig. 6. Comparison of predictive results for key socioeconomic Indicators (Random forest, SARIMA, ridge regression, RBEM)

표 5. 사회경제지표 예측 결과 (MSE)
Table 5. Results of socio-economic indicator (MSE)

socio-economic indicator	RBEM (Proposed)	MEAN (m_1,m_2,m_3)	SARIMA (m_1)	Random forest (m_2)	Ridge regression (m_3)
Household loan	940.2412	5247.7866	515.6971	13214.15	2013.508
House sale price index	2.5799	4.575	1.3342	5.275	7.1162
Price rate	1.6694	2.3887	2.7978	2.8359	1.5323
Unemployment rate	0.1851	0.3685	0.3216	0.1696	0.6142

V. 결 론

본 연구에서는 실시간으로 각 예측모델과 예측모델의 가중치를 최적 조합하는 앙상블 기법과 베이지안 업데이트 기법을 통해 각 모델의 적합도 가중치를 실시간으로 업데이트하는 방법을 활용하는 RBEM 모델을 제안하였다. 사회경제지표는 변수 간 강한 상관관계, 지표 간 선후상관관계의 존재, 시간 변동성 문제로 예측이 어려운 특징을 갖는다. 또한, 기존의 예측모델인 시계열, 머신러닝 모델은 예측시기에 따라 예측 모델의 성능이 상이한 특징을 갖는다. 사회경제지표를 안정적으로 예측하고자 복수의 모델을 앙상블 하는 방법, 모델을 재귀적으로 업데이트하는 방법을 적용하였고, 사례연구에서 예측을 수행할 때 그레인저 인과관계 검정을 통해 선후상관관계를 반영한 예측 모델링을 진행하였다. 예측 결과, RBEM 모델이 단일 시계열 혹은 머신러닝 모델 대비 안정적인 결과를 보였다. 하지만 사회경제지표를 예측함에 있어 급격하게 변화하는 구간과 2020년 코로나 펜데믹으로 사회가 불안정해지는 구간에는 예측의 어려움을 보였고, 이는 데이터 기반 모델의 한계점인 외삽 문제로 사료된다.

최근 공학분야에서 외삽문제를 해결하기 위해 데이터 기반 예측 모델링과 사전지식 기반 예측모델을 결합한 예측 방법이 연구되고 있으며, 사전지식을 고려한 예측모델링은 외삽문제를 해결하는데 용이한 특징을 가진다[24][25]. 따라서 향후에는 경제지식 기반으로 구성된 방정식 모델을 활용하여 외삽 문제를 해결하고자 한다.

References

[1] C. Kim, "Economic Phenomena, Economic Analysis, and Its Statistical Applicability: Focusing on the Developments of Econometrics and Challenging Issues", The Korean Journal of Applied Statistics, Vol. 28, No. 6, pp. 1075-1091, Dec. 2015. https://doi.org/10.5351/KJAS.2015.28.6.1075.

[2] S. H. Han, K. D. Kim, and J. M. Lee, "A Research for the Improvement of Regional Economic Model (KREI-REMO) and Its Linkage with Rural Development Support System (KREI-RDSS)", Korea Rural Economic Institute, pp. 1-103, Nov. 2011.

[3] J. Lee, S.-H. Park, S. Cho, and J.-H. Kim, "Comparison of Models to Forecast Real Estates Index Introducing Machine Learning", Journal of the Architectural Institute of Korea, Vol. 37, No. 1, pp. 191-199, Jan. 2021. https://doi.org/10.5659/JAIK.2021.37.1.191.

[4] D. H. Joo, "Analysis and forecast of the household debt policy effect on the real economy using VARX model," Journal of The Korean Data Analysis Society, Vol. 24, No. 2, pp. 795-810, Apr. 2022. https://doi.org/10.37727/jkdas.2022.24.2.795.

[5] H. S. Hwang, "Hybrid Machine Learning Model for Predicting the Direction of KOSPI Securities," Journal of the Korea Convergence Society, Vol. 12, No. 6, pp. 9-16, Jun. 2021. https://doi.org/10.15207/JKCS.2021.12.6.009.

[6] B. H. Ko, G. C. Kang, and Y. S. Kang, "A Study on the Jeju-do Regional Economic Forecasting Model", Journal of Korean Regional Development Association, Vol. 24, No. 5, pp. 22-39, Dec. 2012.

[7] K. B. Kim, "Regional housing market prediction using machine learning and time series model: Focused on the actual transaction price index", Master's Thesis, Hanyang University Graduate School, Feb. 2020.

[8] C. Y. Im and H. C. Kim, "A Study on the Analysis and Prediction of Housing Mortgage in Deposit Bank Using ARIMA Model", Journal of Korea Institute of Information, Electronics, and Communication Technology, Vol. 12, No. 3, pp. 265-272, Jun. 2019. https://doi.org/10.17661/jkiiect.2019.12.3.265

[9] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the ARIMA model", in 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, pp. 106-112, Mar. 2014. https://doi.org/10.1109/UKSim.2014.67.

[10] C. Cheng et al., "Time series forecasting for nonlinear and non-stationary processes: A review and comparative study", IIE Transactions, Vol. 47, No. 10, pp. 1053-1071, Apr. 2015. https://doi.org/10.1080/0740817X.2014.999180.

[11] K. H. Nam and N. Y. Seong, "Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market", Decision Support Systems, Vol. 117, pp. 100-112, Feb. 2019. https://doi.org/10.1016/j.dss.2018.11.004.

[12] U. C. Buyuksahin and S. Ertekin, "Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition", Neurocomputing, Vol. 361, pp. 151-163, Oct. 2019. https://doi.org/10.1016/jne. com.2019.05.099.

[13] S. Hussain et al., "Structural equation model for evaluating factors affecting quality of social

infrastructure projects", Sustainability, Vol. 10, No. 5, pp. 1415, May 2018. https://doi.org/10.3390/su10051415.

[14] C. H. Wu, C. C. Lu, Y. F. Ma, and R. S. Lu, "A new forecasting framework for bitcoin price with LSTM", in 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 168-175, Nov. 2018.

[15] R. Mulaudzi and R. Ajoodha, "Application of deep learning to forecast the South African unemployment rate: a multivariate approach", in 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1-6, Dec. 2020. https://doi.org/10.1109/CSDE50874.2020.9411581.

[16] S. W. Bae and J. S. Yoo, "Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model", Housing Studies Review, Vol. 26, No. 1, pp. 107-133, Feb. 2018. https://doi.org/10.24957/hsr.20118.26.1.107.

[17] J. W. Kim, "Prediction on the Ratio of Added Value in Industry Using Forecasting Combination based on Machine Learning Method", The Journal of the Korea Contents Association, Vol. 20, No. 12, pp. 49-57, Dec. 2020. https://doi.org/10.5392/JKCA.2020.20.12.049.

[18] A. Ouattara et al., "Scalable Econometrics on Big Data--The Logistic Regression on Spark", arXiv preprint arXiv:2106.10341, Jun. 2021. https://doi.org/10.48550/arXiv.2106.1034.

[19] C. A. Holt and A. M. Smith, "An update on Bayesian updating", Journal of Economic Behavior & Organization, Vol. 69, No. 2, pp. 125-134, Feb. 2009. https://doi.org/10.1016/j.jebo.2007.08.013.

[20] C. Ho. An, "An Empirical Study on Prediction of Household Debts Using a times Series Model", The Society of Convergence Knowledge Transactions, Vol. 6, No. 1, pp. 91-97, Jan. 2018.

[21] E. Y. Choi, S. S. Koh, and H. D. Hawang, "An analysis of the effect of change real estate prices on the bank", Housing Studies Review, Vol. 19, pp. 101-121, Aug. 2011.

[22] J. H. Na and E. S. Kim, "Forecasting unemployment rate using social media information", Journal of the Korea Industrial Information Systems Research, Vol. 18, No. 6, pp. 95-101, Dec. 2013. https://doi.org/10.9723/jksiis.2013.18.6.095.

[23] A. Seth, "Granger causality", Scholarpedia, Vol. 2, No. 7, pp. 1667, 2007. https://doi.org/10.4249/scholarpedia.1667.

[24] T. Amakasu et al., "Conflict-free collective stochastic decision making by orbital angular momentum of photons through quantum interference", Scientific Reports, Vol. 11, No. 1, pp. 21117, Oct. 2021. https://doi.org/10.1038/s41598-02 1-00493-2.

[25] J. Kim et al., "DPM: A novel training method for physics-informed neural networks in extrapolation", in Proc. of the AAAI Conference on Artificial Intelligence, Washington, USA, Vol. 35, No. 9, pp. 8146-8154, May 2021. https://doi.org/10.1609/aaai.v35i9.16992.

저자소개

변 준 현 (Junhyun Byun)

2022년 2월 : 한국공학대학교
IT 경영학과(경영학학사)
2022년 3월 ~ 현재 :
한국공학대학교
스마트팩토리융합학과 석사과정
관심분야 : 머신러닝, 예지보전

민 수 홍 (Suhong Min)

2022년 8월 : 한국공학대학교
산업경영학과(경영학학사)
2022년 9월 ~ 현재 :
한국공학대학교
스마트팩토리융합학과 석사과정
관심분야 : 물리결합 머신러닝

강 지 훈 (Jihoon Kang)

2009년 8월 : 광운대학교
경영학과(경영학학사)
2011년 8월 : 고려대학교
정보경영공학(공학석사)
2015년 2월 : 고려대학교
산업경영공학(공학박사)
2015년 ~ 2018년 : 삼성 SDS
알고리즘 연구소 IoT 랩/데이터분석 랩, Senior Engineer
2018년 ~ 2019년 8월 : 삼성 SDS 빅데이터 분석팀
빅데이터 분석 그룹, Principal Data Scientist
2019년 9월 ~ 현재 : 한국공학대학교 경영학부 조교수
관심분야 : 수리 모델링과 최적화(OR) 기법을 활용한 스마트제조 공정