https://doi.org/10.5392/JKCA.2020.20.12.049

머신러닝 기법 기반의 예측조합 방법을 활용한 산업 부가가치율 예측 연구

Prediction on the Ratio of Added Value in Industry Using Forecasting Combination based on Machine Learning Method

김정우
강릉원주대학교 경제학과

Jeong-Woo Kim(kurtkim@gwnu.ac.kr)

요약

본 연구는 우리나라 수출 분야 산업의 경쟁력을 나타내는 부가가치율을 다양한 머신러닝 기법을 활용하여 예측하였다. 아울러, 예측의 정확성 및 안정성을 높이기 위하여 머신러닝 기법 예측값들에 예측조합 기법을 적용하였다. 특히, 본 연구는 산업별 부가가치율에 영향을 주는 다양한 변수를 고려하기 위하여 재귀적특성제거 방법을 사용하여 주요 변수를 선별한 후 머신러닝 기법에 적용함으로써 예측과정의 효율성을 높였다. 분석 결과, 예측조합 방법에 따른 예측값은 머신러닝 기법 예측값들보다 실제의 산업 부가가치율에 근접한 것으로 나타났다. 또한, 머신러닝 기법의 예측값들이 큰 변동성을 보이는 것과 달리 예측조합 기법은 안정적인 예측값을 나타내었다.

■ 중심어 : | 머신러닝 | 예측 | 예측조합 | 재귀적특성제거 |

Abstract

This study predicts the ratio of added value, which represents the competitiveness of export industries in South Korea, using various machine learning techniques. To enhance the accuracy and stability of prediction, forecast combination technique was applied to predicted values of machine learning techniques. In particular, this study improved the efficiency of the prediction process by selecting key variables out of many variables using recursive feature elimination method and applying them to machine learning techniques. As a result, it was found that the predicted value by the forecast combination method was closer to the actual value than the predicted values of the machine learning techniques. In addition, the forecast combination method showed stable prediction results unlike volatile predicted values by machine learning techniques.

■ keyword : | Machine Learning | Prediction | Forecast Combination | Recursive Feature Elimination |

I. 서 론

주어진 데이터에 대한 학습(Learning) 모델은 두 가지 기준에서 평가될 수 있다. 첫 번째는 특정 학습 모델이 얼마나 주어진 데이터를 잘 설명하고 있는가에 대한 기준이며, 두 번째는 주어진 모델의 예측력과 관련한 기준이다. 전자의 경우는 주어진 샘플 전체를 활용하여 결정계수(R-squared)와 같은 적합도 척도 등으로 모델

접수일자 : 2020년 08월 10일
수정일자 : 2020년 09월 16일
심사완료일 : 2020년 09월 16일
교신저자 : 김정우, e-mail : kurtkim@gwnu.ac.kr

의 설명력을 평가할 수 있다. 반면, 후자의 경우는 예측 대상이 되는 자료가 미래 값으로 주어지지 않는다. 회귀분석 종류의 모델의 경우에는 AIC(Akaike Information Criterion), BIC(Bayesian information criterion), Mallows's Cp 등 상대적으로 다양한 평가 기준이 존재하며, 일반적인 모델에 대해서는 시험오류검증(Test error validation) 방법을 적용해볼 수 있다. 따라서, 학습 모델의 예측력을 평가하는 기준은 다양하며, 하나의 학습 모델에 대한 예측력 평가 기준은 일관적이지 않을 수 있다.

일반적으로 모든 데이터를 잘 설명하고 예측하는 학습 모델은 존재하지 않으며, 이보다는 특정한 데이터에 적합한 학습 모델을 찾아서 적용하는 것이 주어진 데이터를 효율적으로 활용하는 방안이라고 볼 수 있다. 또한, 동일한 종류의 데이터도 사용하는 샘플의 범위 등에 따라서 적합한 학습 모델은 달라질 수 있다. 그러므로, 하나의 모델만을 사용하여 주어진 데이터를 기반으로 예측하고자 하는 것은 예측력의 일관성이라는 부분에서 다소 불리한 전략이라고 볼 수 있다.

예측조합(Forecast combination)은 다수의 모델로부터 얻어진 예측값들을 적절히 조합하여, 하나의 모델에서 얻어진 예측값보다 더 정확하고 안정된 예측을 하기 위해서 제시된 방법이다[1]. 즉, 예측조합은 다수의 예측값들을 토대로 새로운 예측값을 도출하여 이 값을 다시 실제 값에 학습시킨 후 최종적인 예측값을 산출하는 방법이다. 그러므로 이 방법은 일관적이고 정확한 예측을 하는데 유용하게 쓰일 수 있다. 또한, 예측조합 방법은 기본 적으로 가중치 평균이므로 서로 이질적인 학습 모델들을 사용할수록 각 학습 모델의 편의나 분산을 완화할 수 있다. 그러므로, 본 연구에서는 다양한 학습 모델 방법을 고려하기 위하여 다수의 머신러닝 기법들에 예측조합 방법을 적용하였다. 아울러, 본 연구에서는 우리나라의 주요 수출분야인 반도체, 통신기기, 자동차, 조선 산업의 연도별 부가가치율과 이와 관련한 경제변수들을 사용하여 머신러닝 기법을 적용한 예측조합 방법의 예측성능을 점검해보았다. 아울러 기업의 영업이익, 인건비 등을 포함한 부가가치액이 매출에서 차지하는 비중인 부가가치율을 예측 대상 변수로 설정하여, 본 연구는 우리나라의 실질적인 산업별 수출 경쟁력을 고려하고자 하였다.

머신러닝 기법을 산업별 부가가치율에 적용한 연구한 사례는 많지 않은 편이다. 다만, 부가가치와 관련한 국내총생산(Gross domestic product, GDP) 자료를 활용한 머신러닝 연구는 상대적으로 많은 편이다. Bae et al. (2018)은 부동산 가격 지수 예측 연구에서 서포트벡터머신(Support vector machine, SVM), 랜덤포레스트(Random forest) 등의 머신러닝 기법을 GDP, GDP 성장률, 소비자 물가지수 등의 다양한 변수들에 적용하였다[2]. Choi et al. (2017)은 GDP, 금리 등의 다양한 경제변수를 사용하여 SVM, 신경망(Neural network) 모형, 랜덤포레스트 등의 머신러닝 기법을 통한 기업의 부도예측 연구를 수행하였다[3]. Lee et al. (2015)은 매출 등의 기업실적을 예측하는 변수 생성을 위하여 주성분분석(Principal component analysis, PCA)과 신경망(Neural network) 모형을 활용하였는데, 특히, 이 연구에서는 PCA를 기반으로 사용된 변수들을 인위적으로 제거하기 보다는 다중공선성을 감소시켜 예측 변수를 선별해내었다[4].

데이터의 일부만을 이용하는 k-최근접 이웃법(k-Nearest Neighbor, kNN), k-평균 군집화(k-means Clustering) 등도 예측 연구에서 사용되는 빈도가 높은 머신러닝 기법들이다. Guegan et al. (2009)는 kNN을 활용한 회귀분석을 GDP 예측에 적용하였으며[5], Ferrara et al (2009)은 GDP 예측에 k-평균 군집화 기법을 적용하였다[6]. 이와 같은 머신러닝 기법은, 복잡한 추세를 보이는 데이터의 예측을 위하여 필요한 범위만을 사용한다는 점에서 효율적인 머신러닝 예측 기법이라고 할 수 있다.

일반적인 회귀분석에서는 관측치의 개수가 변수 개수보다 많아야 회귀계수가 추정가능하다. 하지만, 능선회귀분석(Ridge regression), LASSO(Least Absolute Shrinkage and Selection Operator) 등의 기법들은 관측치가 부족한 경우에도 사용이 가능하다. Song (2015)은 1인당 GDP, 환율 등을 사용하여 금리를 예측하는 연구에서 LASSO를 사용하였다[7]. Fokin et al. (2019)은 벡터자기회귀모형(Vector autoregression)에 LASSO를 결합한 기법으로 GDP를 예측하는 연구를 수행하여, 다변량시계열 데이터에서의 머신러닝 기법의 예측 성능을 검증하였다[8].

본 연구는 상기의 다양한 머신러닝 기법으로 산업별 부가가치율을 예측하고, 더 나아가서 예측조합 기법을 통한 예측 성능이 개선되는지를 알아보고자 하였다. 특히, 부가가치율에 영향을 주는 요소가 다양하다는 점을 고려하여 관측치보다 많은 독립변수를 사용하였으며, 이러한 상황에 적합한 머신러닝 기법들을 사용하였다.

본론에서는 자료 전처리 방법과 본 연구에서 사용된 머신러닝 기법 및 예측조합 등에 기술하고, 산업별 부가가치율 예측 결과를 비교, 검토한다. 결론에서는 본 연구의 의의, 향후 연구주제 등이 논의될 것이다.

II. 본 론

1. 자료 전처리

본 연구에서는 산업별 부가가치율을 예측하기 위하여 필요한 26개의 독립변수가 사용되었다. 부가가치는 총생산물에서 원재료, 감가상각 등을 제외한 것으로 기본적으로 생산요소에 분배되는 소득과 같다. 그러므로, 자본, 노동 등의 공급측면 요소와 민간소비, 정부지출 등 수요측면 요소 등의 국내변수를 포함하였다. 아울러, 본 연구에서는 수출 산업의 부가가치율을 다루고 있으므로 해당 산업 생산물의 해외수요와 가격에 영향을 주는 환율, 유가, 세계물가 등의 국제변수도 고려되었다.

총 관측치수는 국민계정을 측정하는 기준연도인 2005년부터 2018년까지 총 14개로, 고려되는 독립변수의 개수보다 적다. 이에 따라, 주어진 데이터를 가지고 회귀분석 등을 시도하게 되면 과소결정(Underdetermined) 문제에 귀착하게 된다. 이에 따라, 본 연구는 변수 선별을 위한 자료 전처리 및 적절한 머신러닝 기법을 사용하여 산업별 부가가치율의 원활한 예측을 도모하였다.

또한, 서로 관련된 많은 독립변수를 사용하는 경우 다중공선성 문제가 발생하는 경우가 있어 학습 모델의 예측성능이 저하될 수 있다. [그림 1]은 본 연구에서 사용된 변수들(반도체) 간의 상관계수를 Heatmap으로 나타낸 것이며, 모눈의 색이 진할수록 높은 양(+) 또는 음(-)의 상관관계를 나타내고 있다. 대부분의 변수들이 짙은 색의 모눈을 나타내고 있으므로 본 연구에서 사용되는 독립변수들은 밀접하게 연관된 경우가 많은 것으로 파악된다. 이에 따라, 본 연구에서 머신러닝 모델을 데이터에 적용하기 이전 단계에서, 독립변수들을 선별하는 작업이 필요할 것으로 판단된다.

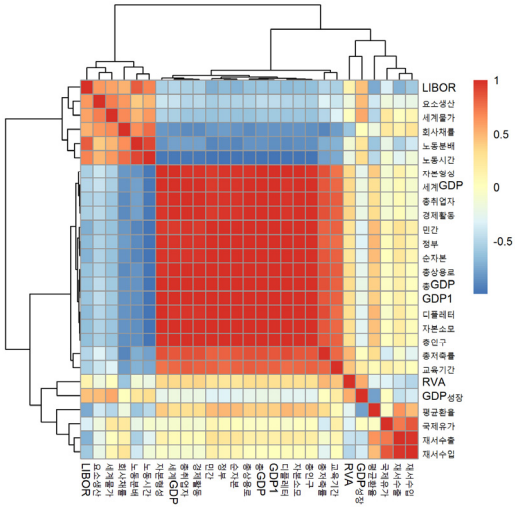


그림 1. 상관계수 Heatmap

재귀적특성제거(Recursive feature elimination, RFE)는 이러한 독립변수 선별에 사용되는 기법들 중 하나로써, 특히 관측치가 적은 경우에 적용되는 방법이다[9]. RFE는 우선 모든 독립변수를 종속변수에 적합(Fitting)시킨 후, 높은 종속변수 설명력을 보일 때까지 독립변수들을 제거해나가는 후진제거법(Backward selection) 방식을 취하고 있다. 본 연구에서 데이터 개수보다 많은 독립변수의 개수를 고려하여 머신러닝 방법 중 하나인 랜덤포레스트(Random forest)를 독립변수를 종속변수에 적합시키는 과정에 사용하였다. RFE의 방법의 알고리즘은 [그림 2]와 같다[10].

Algorithm: Recursive feature elimination

1 Tune/train the model oil the training set using all predictors
2 Calculate model performance
3 Calculate variable importance or rankings
4 **for** Each subset size S_i, i = 1 ... S **do**
5 | Keep the S_i most important variables
6 | [Optional] Pre-process the data
7 | Tune/train the model on the training set using S_i predictors
8 | Calculate model performance
9 | [Optional] Recalculate the rankings for each predictor
10 **end**
11 Calculate the performance profile over the S_i
12 Determine the appropriate number of predictors
13 Use the model corresponding to the optimal S_i

그림 2. RFE 알고리즘

2. 머신러닝 기법

예측을 위하여 데이터 분석을 하는 경우에 주어진 데이터를 모두 활용하여 주어진 데이터 내에서의 설명력을 높이는 접근 방법은 과적합(Overfitting) 문제를 초래할 수 있다. 과적합 문제가 발생하면 사용된 학습 모델이 주어진 자료는 잘 설명하는 반면, 주어진 자료를 벗어나는 예측값들에 대한 예측력은 저하될 수 있다. 또한, 과적합 문제는 학습 모델 복잡도(복잡한 함수 형태 등)가 증가할수록 커지는 경향을 나타내는 것으로 알려져 있다[11].

학습 모델의 예측력을 높이기 위하여 모델 복잡도의 문제를 완화하는 방법 중 대표적인 것이 정규화(Regularization) 방법이다. 일반적으로 정규화는 회귀계수 등의 추정된 파라미터의 크기를 감소시키는 방향으로 학습 모델을 최적화하여 모델의 예측 성능을 높이게 된다. 대표적인 모델 정규화 기법으로는 Ridge regression, LASSO 등과 같은 기법이 있다. 특히, LASSO는 변수의 회귀계수를 0으로 추정할 수 있으므로, 모델의 복잡도를 감소시킬 수 있으므로 관측치 개수보다 많은 변수들을 사용한 본 연구에서 적합한 머신러닝 기법 중 하나라고 할 수 있다.

LASSO가 기본적으로 하나의 함수 모형을 설정하고 추정하는데 반면, 서포트벡터머신은 주어진 데이터의 예측값들의 거리(Margin)를 최대한 구분하는 방식으로만 단순 선형모형을 추정하기 때문에 서포트벡터머신에 기반한 서포트벡터회귀 또한 모델 복잡도를 낮추는 방법으로 사용될 수 있다.

모델 복잡도 문제를 완화하고 예측력을 제고시킬 수 있는 또 다른 방법 중 하나로써 주어진 데이터 중 일부를 사용하는 방법이 있다. 이러한 접근법은 주어진 데이터를 모두 사용하지 않으므로, 전체 데이터 적합을 위해 필요한 복잡한 함수형태가 필요하지 않으므로 모델 복잡도로 인한 과적합 문제를 완화하는데 도움이 될 수 있다. 데이터의 일부분만을 사용하는 방법에는 대표적으로 k-최근접 이웃법과 k-평균 군집화 등이 있다. 이러한 방법들은 주어진 자료에서 우리가 예측하고자 하는 값들과 인접한 데이터를 선별하고, 이 데이터에 단순한 형태의 함수를 적용함으로써 모델 복잡도를 낮추고, 예측력을 높이는 데에 유용한 방법으로 쓰일 수 있다.

아울러, 본 연구에서는 비교적 많은 양의 독립변수가 사용되므로 다중공선성 문제가 발생할 수 있다. 주성분분석(Principal component analysis, PCA)은 데이터의 변수들을 선형변환하여 서로 직교하는 새로운 변수들을 생성하는 방법이다. 주성분회귀(Principal component regression, PCR)은 이 새로운 변수들을 회귀분석에 적용한 것으로써 다중공선성 문제를 감소시킬 수 있는 방법이다.

신경망 모형은 변수들 간의 가중치를 다르게 설정하여 새로운 변수들을 생성하고, 종속변수에 적합시키는 방법이다. 신경망 모형은 새로운 변수들을 종속변수에 적합시키는 과정에서 과적합 문제가 발생할 수 있으나, 예측값과 실제값의 차이를 역전파(Backpropagation) 과정을 통하여 가중치를 갱신하므로 예측력을 높일 수 있는 방법이다.

3. 예측조합 방법

본 연구에서 쓰인 관측치 개수는 2005년부터 2018년까지 총 14개로 독립변수의 개수보다 적으므로 예측값들의 분산이 클 수가 있다. 예측조합 기법을 사용하면 다수의 예측값들을 적절히 조합하여 최종 예측값을 도출함으로써 이러한 문제를 완화할 수 있다. 예측조합 기법의 기본적인 수식은 아래와 같다.

$$f_c = \sum_{i=1}^{m} \omega_i f_i \tag{1}$$

여기서, f_c는 최종예측값, m은 예측 모델의 개수, w_i는 각 예측 모델의 예측값에 부여되는 가중치, f_i는 각 예측 모델의 예측값임.

예측조합 기법 활용 시 중요한 사항은 어떠한 예측값들을 또는 얼마나 많은 예측값들을 예측조합 기법에 포함시켜야 하는가에 대한 문제가 예측조합 기법의 성능에 영향을 줄 수 있다는 점이다. 이상적인 전략은 우수한 예측성능을 가지는 상이한 예측 기법들를 최대한 많이 예측조합 기법에 포함하는 것이다. 하지만, 데이터의 특성, 시공간 등의 차이로 언제나 우수한 예측성능을

지니는 예측기법을 찾아내거나 개발하여 예측조합 기법에 포함하는 것은 용이하지 않다. 그러므로 서로 다른 접근방식의 예측기법들을 예측조합 기법에 포함하여 예측기법들 간의 분산을 낮추어 예측성능을 개선하는 전략이 절충적 방안이 될 수 있다. 이에 따라 본 연구에서는 상기 언급한 바와 같이 정규화 방식, 데이터의 일부를 사용하는 방식, 딥러닝 방식의 머신러닝 기법들을 예측조합 기법에 포함하여 예측조합 기법의 장점을 최대화하고자 하였다.

아울러, 데이터의 특성, 시공간 등을 고려하여 적합한 예측조합 기법을 사용하는 것도 예측성능을 높이는 데 중요하다. 우선, 각 예측값에 등일하게 가중치를 부여하는 예측조합 기법이 있다. 이 예측조합 기법은 간단하지만 실용적인 측면에서 널리 활용되는 예측조합 기법이다[12]. 또한, 각 예측값들을 실제 값에 회귀하여 가중치를 결정하는 방법[13]이 있는데, 각 예측값들을 독립변수로 간주한다는 점에서 직관적인 방법이라고 볼 수 있다.

각 예측값에 동일한 가중치를 부여하는 방식은 예측기법들 중 높은 예측성능의 예측기법에 높은 가중치를 부여하지 못한다는 단점이 있다. 회귀방식에 따른 가중치 부여 방법은 회귀분석의 일반적인 단점과 마찬가지로 이상치(Outlier)에 민감하여 예측의 안정성이 저하될 수 있는 위험을 내포하고 있다.

한편, 예측값들의 평균제곱오차 (Mean squared error, MSE)에 반비례하는 가중치를 부여하는 방법도 널리 활용되고 있는 예측조합 기법이다[1]. MSE에 반비례하는 가중치는 아래와 같이 결정된다.

$$\omega_i = \frac{\hat{\sigma}_i^{-2}}{\sum_{j=1}^{m} \hat{\sigma}_j^{-2}} \quad (2)$$

여기서, $\hat{\sigma}_i^{-2}$는 i번째 예측기법의 MSE임.

MSE 기반의 예측조합 기법은 각 예측값들의 예측오류를 기준으로 가중치를 부여하므로 예측성능이 높은 예측기법에 높은 가중치를 부여할 수 있으며, 각 예측기법의 MSE를 개별적으로 고려하므로 회귀방식의 가중치 부여 방법보다 이상치에 덜 민감할 수 있다. 따라서, 본 연구에서는 각 예측기법의 예측값들의 MSE에 반비례하는 가중치를 부여하는 방식을 사용하기로 한다.

4. 분석 결과

본 연구에서의 예측 절차는 다음과 같다. 우선, 주어진 관측치 수보다 많은 변수를 고려하여 RFE로 자료들을 전처리하여 주요 변수들을 선별하였다. 선별된 변수를 사용하여 6개의 머신러닝 기법을 적용하여 각 예측값을 도출하였으며, 이 예측값들에 예측조합 기법을 적용하여 각 산업별 부가가치율을 예측하였다. 예측에 사용된 기간은 2005년부터 2017년까지이며 이 기간의 자료를 토대로 2013년부터 2018년까지의 예측값들을 구하여 각 머신러닝 기법과 예측조합 기법의 예측결과들을 산출하였다.

[그림 3]과 같이 우리나라의 주요 수출 품목 중 부가가치율이 가장 높은 산업은 반도체이며 통신기기가 그 뒤를 잇고 있다. 통신기기의 경우 중국 등의 후발업체들의 휴대폰 시장 확대로 부가가치율이 점점 낮아지고 있는 것으로 파악된다.

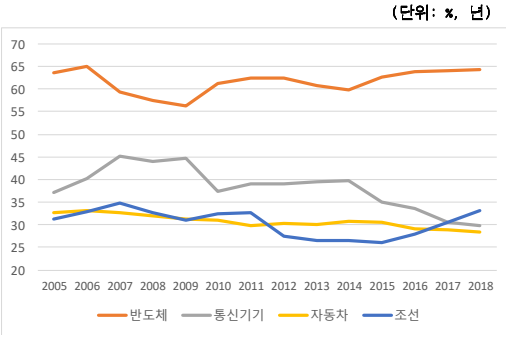


그림 3. 산업별 부가가치율 추이

또한, 기술통계량에서도 통신기기의 부가가치율은 평균값과 중간값이 반도체보다 낮은 수준이나, 표준편차가 상당히 높은 편으로 나타나 통신기기의 부가가치율은 시장 환경에 따라 변동성이 큰 것으로 파악된다. 이러한 변동성이 큰 데이터 특성은 예측 결과에도 영향을 줄 수 있다고 볼 수 있다.

표 1. 기술통계량 **(단위: %)**

산업	MEAN	STD	MIN	0.25	0.5	0.75	MAX
반도체	61.64	2.65	56.29	60.11	62.42	63.81	65.02
통신기기	38.18	4.76	29.72	35.58	39.00	40.17	45.08
자동차	30.72	1.51	28.33	29.81	30.65	31.68	33.23
조선	30.40	2.95	25.98	27.62	31.15	32.69	34.73

MEAN: 평균, STD: 표준편차

예측 결과를 살펴보기 전에, RFE 방법에 따라 선택된 변수들에 대하여 알아보고자 한다. [표 2]에서와 같이 각 연도별로 선택된 변수들은 차이를 보이고 있다. 한편, 특정 변수들은 여러 연도에 걸쳐 중요성이 높은 변수로 RFE에 의해 공통적으로 선별된 것이 발견된다. 예를 들어, 자동차 산업에서는 총 상용근로자수(총상용로)가 대부분의 연도에 걸쳐 중요한 변수로 채택되어, 자동차의 조립 공정이 노동집약적이라는 것과 연관이 있다고 하겠다. 이와 같은 이론적 기반의 변수 관계에 대한 설명은 충분한 양의 데이터 개수가 확보되어야만 통계적 유의성에 의해 뒷받침될 수 있다. 하지만, 본 연구는 변수간의 이론적인 관계를 증명하기보다는 주어진 데이터를 기반으로 예측을 목적으로 하고 있으므로, RFE에 의해 선별된 공통된 변수들을 머신러닝 기법에 우선 사용하기로 한다.

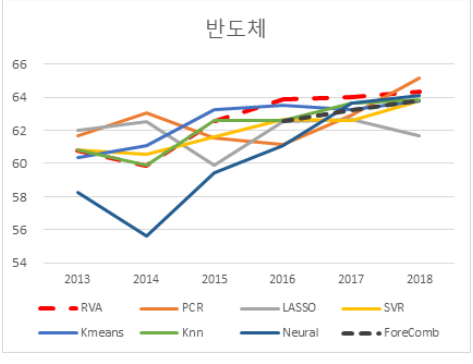
표 2. RFE 방법에 의한 선택변수

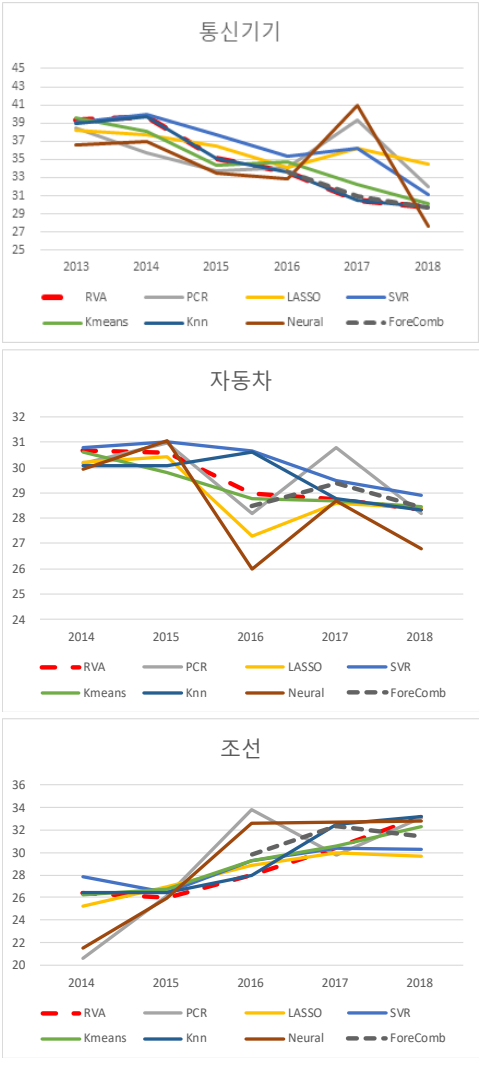
연도	반도체	통신기기	자동차	조선
2013	회사채률, 총상용로, 노동분배, GDP1	회사채률, 총저축률, 노동분배, 총GDP, 총상용로, 디플레터, GDP1, 요소생산	총상용로, 경제활동, 자본형성, 자본소모, LIBOR, 총인구, 노동분배, 노동시간	요소생산, 회사채률, 총상용로, 순자본, 세계물가, 노동시간, 정부, 세계GDP, 자본형성, 총인구, 민간, 총취업자, 총GDP, 교육기간, 디플레터, 경제활동
2014	회사채률, 요소생산, 총상용로, 민간, GDP성장, GDP1, 자본소모, 자본형성	회사채률, 총저축률, GDP성장, 노동분배	노동시간, 디플레터, LIBOR, 노동분배, 순자본, 자본소모, 민간, 자본형성	요소생산, 총상용로, 노동시간, 회사채률
2015	회사채률, 재서수입, 재서수출, 국제유가, GDP1, GDP성장, 세계GDP, 노동분배	총저축률, 회사채률, 국제유가, 총인구	총GDP, 자본소모, 총인구, 디플레터, 총상용로, 자본형성, 정부, 민간	회사채률, 총상용로, 총인구, 순자본, 자본소모, 노동시간, 디플레터, 교육기간, 총GDP, 세계GDP, 경제활동, 총취업자, 민간, GDP1, 요소생산, 자본형성, 정부, 세계물가, GDP성장, 총저축률, LIBOR, 국제유가, 평균환율, 노동분배, 재서수출, 재서수입
2016	회사채률, 재서수입, 재서수출, GDP성장	총저축률, 회사채률, 국제유가, 세계GDP	총상용로, 순자본, 정부, 총GDP, 경제활동, 자본형성, 총인구, 디플레터	요소생산, 회사채률, 민간, 세계물가, 자본형성, 총인구, 정부, 디플레터, 총상용로, 교육기간, 순자본, 노동시간, GDP성장, 세계GDP, 총취업자, 자본소모
2017	회사채률, 재서수입, 요소생산, 재서수출, 국제유가, GDP1, 총저축률, GDP성장, 디플레터, 자본소모, 경제활동, 총취업자, 총상용로, 노동분배, 총인구, 총GDP	총저축률, 회사채률, 총GDP, 디플레터, GDP1, 자본소모, 총인구, 자본형성	민간, 노동시간, 총인구, 총GDP, 자본소모, 순자본, 자본형성, 노동분배, 총상용로, 경제활동, 세계GDP, GDP1, 디플레터, 총취업자, 정부, 총저축률	요소생산, 교육기간, 세계물가, LIBOR, 회사채률, 국제유가, GDP성장, 총상용로

[그림 4]는 각 산업별로 머신러닝 기법과 예측조합 기법에 의한 예측추세를 보여주고 있다. 붉은색 쇄선이 실제 산업별 부가가치율이며 검은색 쇄선이 예측조합 기법에 따른 예측값이다. 예측조합 기법에 따른 예측값은 이전의 머신러닝 기법 예측값들을 기반으로 얻을 수 있으므로, 2016년부터 2018년까지의 예측값만 제시되었다.

[그림 4]에서와 같이 산업별 및 연도별로 우수한 예측 성능을 보이는 머신러닝 기법은 일관적으로 관찰되지 않는다. 반면에, 예측조합에 의한 예측값은 대부분 부가가치율에 근접한 모습을 보여주고 있으며, 특히 분산이 적은 것으로 관찰된다. 이것은 예측조합 기법이 다양한 머신러닝 기법의 가중치 평균이라는 점에 기인하는 것으로, 안정적인 예측값을 얻는데 예측조합 기법을 사용하는 것이 유리하다는 것을 보여준다고 하겠다.

(단위:%, 년)





RVA: 산업 부가가치율, ForeComb: 예측조합

그림 4. 예측결과 비교

아울러, 각 기법들의 예측성능을 보다 정확하게 비교하기 위하여 예측오차의 평균절대비오차(Mean Absolute Percentage Error, MAPE)를 구하였다. MAPE의 수식은 다음과 같다.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - f_i|}{y_i} \times 100 \qquad (3)$$

여기서, N은 데이터 개수, y_i는 실제 GDP값, f_i은 예측값임.

각 머신러닝 기법은 산업에 따라 상이한 예측성능을 나타내고 있다. 예를 들어, LASSO의 경우에는 반도체와 자동차 분야에서 평균적으로 양호한 예측성능을 보이고 있으나, 통신기기와 조선업 분야에서는 높은 MAPE를 나타내고 있다. 아울러, 통신기기의 경우 머신러닝 기법들의 예측 성능이 높은 편이 아닌데, 이것은 기술통계량에서 언급된 바와 같이 통신기기 부가기치율의 큰 변동성에 기인한다고 볼 수 있다.

한편, 예측조합 기법의 MAPE는 다른 머신러닝 기법들의 MAPE에 비해 낮은 수준으로 나타났다. 특히, 각 머신러닝 기법의 연도별 MAPE 값들의 표준편차 비해 예측조합 기법의 MAPE의 표준편차는 적게 나타나 안정적인 예측성능을 보여준다고 볼 수 있다. 조선업의 경우에서 예측조합 기법의 예측성능은 다소 낮은 수준을 보이고 있는데, 이것은 다른 머신러닝 기법의 MAPE가 전체적으로 높은 것에도 기인하나, MSE 기반의 가중치 평균의 예측조합 기법의 특징이라고도 볼 수 있다. 예를 들어, 2018년 PCR의 MAPE는 상당히 낮은 수준임에도 불구하고 이전 MAPE 값들이 높은 수준이므로 2018년 예측조합의 예측값에는 2018년 PCR의 예측값의 가중치가 낮게 된다. 이에 따라, 2018년 예측조합의 MAPE가 중간 수준으로 나타나게 되는 것이다. PCR 기법은 모든 산업에서 MAPE의 표준편차가 높은 수준으로 관찰되는데, 이것은 예측성능이 연도에 따라 차이가 크다는 것을 의미한다. 일반적으로 위험기피적인 예측을 가정할 시에는 이러한 변동이 큰 예측값들의 위험성을 완화시켜주는 것이 예측조합 기법의 이점이라고 볼 수 있다.

표 3. MAPE 비교 **(단위: 년, %)**

연도	PCR	LASSO	SVR	Kmeans	Knn	Neural	Fore Comb
	반도체						
2013	3.02	3.51	1.55	0.77	1.54	2.76	-
2014	0.75	0.06	3.24	2.37	4.28	11.07	-
2015	3.65	6.24	3.56	1.03	2.05	6.99	-
2016	4.46	2.30	2.13	0.83	2.27	4.60	2.04
2017	2.16	2.54	2.67	1.69	1.10	1.07	1.26
2018	3.02	2.53	0.79	0.85	0.99	1.33	0.84
MEA	2.84	2.86	2.32	1.26	2.04	4.64	1.38

N							
STD	1.28	2.01	1.05	0.64	1.21	3.85	0.61
	통신기기						
2013	3.13	3.69	1.41	0.40	1.79	7.74	-
2014	1.99	7.59	13.85	8.45	13.14	5.26	-
2015	0.41	8.28	12.18	2.12	4.25	0.44	-
2016	11.73	11.70	15.97	13.89	10.36	7.57	0.23
2017	32.20	21.86	21.79	8.49	2.62	38.01	1.29
2018	5.42	13.59	2.63	0.50	2.04	8.69	0.11
MEAN	9.15	11.12	11.31	5.64	5.70	11.29	0.54
STD	11.96	6.29	7.90	5.48	4.84	13.42	0.65
	자동차						
2013	1.87	2.36	1.15	0.98	1.14	3.10	-
2014	1.45	1.27	0.69	0.00	1.67	2.21	-
2015	6.79	4.97	6.93	2.78	3.79	7.10	-
2016	1.92	5.10	6.65	0.00	6.40	9.64	1.74
2017	8.73	0.87	4.16	1.28	1.52	1.30	2.22
2018	0.04	0.83	2.54	1.00	0.47	5.05	0.30
MEAN	3.47	2.57	3.69	1.01	2.50	4.73	1.42
STD	3.45	1.99	2.69	1.02	2.21	3.18	1.00
	조선						
2013	5.27	13.09	9.76	10.05	0.08	12.56	-
2014	20.79	2.81	7.32	0.87	1.73	17.06	-
2015	6.66	3.74	5.57	4.27	5.54	7.19	-
2016	10.98	5.33	3.91	4.05	8.11	7.13	6.56
2017	10.09	9.71	8.28	7.91	2.02	1.46	6.24
2018	0.05	10.37	8.62	2.66	0.08	1.10	5.22
MEAN	8.97	7.51	7.24	4.97	2.93	7.75	6.01
STD	6.98	4.13	2.16	3.40	3.23	6.23	0.70

MEAN: MAPE의 평균, STD: MAPE의 표준편차

Ⅲ. 결 론

본 연구에서는 다양한 머신러닝 기법으로 우리나라의 주요 수출분야인 반도체, 통신기기, 자동차, 조선업의 부가가치율을 예측하여 각 기법들의 예측 성능을 비교해보았다.

아울러, 많은 실용적 문제에서 직면하는 독립변수 개수보다 적은 관측치가 주어지는 상황을 고려하여 재귀적특성제거 기법으로 자료 전처리를 수행하였다. 이에 따라 각 산업별 또는 연도별로 부가가치율에 영향을 주는 중요한 변수들을 알아볼 수 있었다.

전처리된 자료를 활용하여 머신러닝 기법을 통한 예측값들을 예측조합 기법에 적용하여 예측 성능의 개선 여부를 검증해보았다. 예측조합 기법은 여러 가지 예측값들을 적절하게 조합하여 새로운 예측값들을 도출하는 방식이다. 그러므로, 상이한 예측값들을 조합할수록 예측조합의 장점이 극대화된다고 볼 수 있다. 이에 따라, 본 연구에서는 머신러닝 기법 중 정규화 방식, 데이터의 일부만을 사용하는 방식, 딥러닝 방식의 머신러닝 기법들을 예측조합 기법에 포함하였다.

본 연구에서 사용된 예측조합 기법은 MSE 기반의 가중치 평균을 사용한 방식이다. 이 예측조합 방식은 예측 오류가 적은 예측 기법들에 큰 가중치를 부여하는 방식으로 기존의 동일 가중치 부여방식이나 회귀분석 방식의 가중치 부여방식보다 예측의 정확성과 안정성이 높다고 할 수 있다.

아울러, 이전의 예측값들을 어느 정도로 사용해야 하는지가 쟁점이 될 수가 있다. 데이터의 종속변수가 불규칙한 추세를 보인다면, 비교적 최근의 예측값들만을 고려한 예측조합 기법을 사용하는 것이 유리할 수 있다. 하지만, 이러한 경우 예측 결과의 분산이 커져서 불안정한 예측이 될 수 있는 단점이 있다. 즉, 예측조합 기법에서도 편향-분산 트레이드 오프(Bias-variance tradeoff) 문제가 개입되므로, 예측조합 기법에 사용되는 예측값들의 범위를 데이터의 추세에 따라 결정하는 새로운 방법이 향후 연구에서 요구된다.

참 고 문 헌

[1] J. M. Bates and C. W. J. Granger, "The Combination of Forecasts," Journal of the Operational Research Society, Vol.20, No.4, pp.451-468, 1969.

[2] 배성완, 유정석, "머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측," 주택연구, 제26권, pp.107-133, 2018.

[3] 오세경, 최정원, 장재원, "빅데이터를 이용한 딥러닝 기반의 기업 부도예측 연구," KIF working paper, 제8호, pp.1-113, 2017.

[4] 이준혁, 김갑조, 박상성, 장동식, "PCA 를 활용한 기업실적 예측변수 생성," Journal of Korean Institute of Intelligent Systems, 제25권, 제2호, pp.191-196, 2015.

[5] D. Guegan and P. Rakotomarolahy, "The multivariate k-nearest neighbor model for dependent variables: one-sided estimation and

forecasting," Documents de travail du Centre d'Economie de la Sorbonne 09050, Universite Pantheon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne, 2009.

[6] L. Ferrara, D. Guegan, and P. Rakotomarolahy, "GDP nowcasting with ragged-edge data: a semi-parametric modeling," Journal of Forecasting, Vol.29, No.1-2, pp.186-199, 2010.

[7] 송상윤, "예대금리차 결정요인 모형의 예측력 비교 연구-Ridge, LASSO 및 Elastic Net 방법론을 중심으로," 금융지식연구, 제13권, 제3호, pp.41-65, 2015.

[8] N. Fokin and A. Polbin, "Forecasting Russia's Key Macroeconomic Indicators with the VAR-LASSO Model," Russian Journal of Money and Finance, Vol.78, No.2, pp.67-93, 2019.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine learning, Vol.46, No.1-3, pp.389-422, 2002.

[10] K. E. Rao and G. A. Rao, "Ensemble learning with recursive feature elimination integrated software effort estimation: a novel approach," EVolutionary Intelligence, pp.1-12, 2020.

[11] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: data mining, inference, and prediction," Springer Science & Business Media, p.38, 2009.

[12] R. T. Clemen, "Combining forecasts: A review and annotated bibliography," International journal of forecasting, Vol.5, No.4, pp.559-583, 1989.

[13] C. W. Granger and R. Ramanathan, "Improved methods of combining forecasts," Journal of forecasting, Vol.3, No.2, pp.197-204, 1984.

저 자 소 개

김 정 우(Jeong-Woo Kim) 정회원

- 2005년 8월 : 고려대학교 경제학사
- 2012년 8월 : 연세대학교 경제학 석사
- 2018년 2월 : 연세대학교 경제학 박사
- 2020년 3월 ~ 현재 : 강릉원주대학교 경제학과 교수

〈관심분야〉 : 계량경제학, 머신러닝