

가우시안 커널 네트워크를 이용한 원유가격예측

김동민*, 신성국*, 길이만**

Crude oil price prediction based on Gaussian Kernel Function Networks

Dongmin Kim*, Sung Kuk Shyn*, and Rhee Man Kil**

요 약

원유가격은 그 중요성에도 불구하고 높은 변동성으로 인하여 가격 예측의 어려움이 많다. 이러한 관점에서 본 논문은 원유시장의 변동성에 알맞은 모델로 가우시안 커널 네트워크를 이용하여 원유가격예측을 진행한다. 가우시안 커널 네트워크(GKFN)는 커널의 개수, 데이터 간격 등 연구자의 직관에 의존한 다양한 실험변수들을 체계적으로 설정하고, 가우시안 함수를 바탕으로 학습을 최적화 시키는 장점을 가지고 있다. 본 논문은 일간 분석, 주간 분석, 월간 분석으로 연구를 진행하였다. 각 데이터별로 GKFN를 구성하고 데이터 부족으로 인한 과적합 및 정확도 저하의 현상은 일 데이터의 재가공을 통해 정확도를 향상시켰다. GKFN 모델은 여러 확장 기술을 통해 단기에서 장기 분석으로 넘어가면서 정확도 손실을 바로 잡아 어떤 데이터로도 일반화할 수 있는 모델로 구성할 수 있다.

Abstract

Although the importance of crude oil price is getting higher, its prediction is a very difficult task due to the highly volatile nature in oil market. In this point of view, this paper conducted crude oil price prediction task based on Gaussian Kernel Function Networks. Gaussian Kernel Function Networks(GKFN) can systematically set its hyperparameters that were previously set based on researchers' intuition such as kernel numbers, intervals of dataset, etc. Furthermore, it maximizes its learning capacity by setting models' kernel based on Gaussian function. This paper analyzes daily, weekly, and monthly fluctuations of crude oil price respectively. GKFN is established for each dataset and the problems of overfitting and low-precision are resolved by rearranging training data. By several extended techniques, GKFN remedies the loss of accuracy as the analysis expands from short-term to long-term, which composes the well-generalized model of any data series.

Key words

crude oil price prediction, Gaussian Kernel Function Networks, smoothness test

* 성균관대학교(Sungkyunkwan University) 글로벌경제학과/컴퓨터공학과 학부과정

** 성균관대학교(Sungkyunkwan University) 소프트웨어대학 교수

1. 서론

원유는 세계 각국의 산업에 있어 필수불가결한 재화다. 원유의 가격에 따라 산업, 특히 운송, 제조 산업의 흥망이 결정되고 원유의 가격은 세계 경제 흐름에 따라 변하며 현재의 경제 상황을 담아내고 있는 지수이기도 하다. 그렇기에 원유시장의 변동성은 각 산업에 큰 영향력을 주고, 크게는 국가경제의 기반을 흔드는 주요 원인이 되기도 한다. 특히, 2020년 4월 원유 파생상품 매수포지션의 실현으로 원유가격이 마이너스를 기록하는 기현상이 일어나며 그 불확실성 또한 증대되고 있으며, 이에 실생활 및 각종 산업에 큰 영향을 미치며 국가경제에 미치는 영향은 막대한 것으로 드러나고 있다.

하지만, 원유가격은 타 재화와 다르게 예측의 어려움이 크다. 이는 원유시장이 가진 특수한 시장 구조에 기인한다.

우선, 원유시장은 공급자가 한정되어 있다는 특징을 가진다. 원유시장은 소수의 공급자가 지배하는 시장 구조로 인해 정치적인 요소에 영향을 많이 받으며, 사업 공급자 간의 상호 결정 역시 중요한 요소이다. 원유 시장은 또한 전반적인 경제상황, 지역적 특성, 국가 상황에 많은 영향을 받는다. 이러한 특성들은 예측할 수 없는 위험(unpredictable risk)으로 수치적으로 계산할 수 없는 변동성에 해당한다. 이와 같은 특성으로 인해, 회귀분석 기법과 같은 전통적인 접근 방식은 다양한 변수를 한꺼번에 반영하여 담기에 무리가 있다.

1.A. 선행 연구

원유가격이 가지는 중요성으로 인해 원유 가격을 분석하고 예측하는 다양한 기법들이 등장하였다. 전통적인 기법들로 시계열 데이터를 가지고 Random Walk 모델 방식을 통한 예측[4], ARIMA 모델을 기반으로 한 예측[5] 방법이 소개되었다. 하지만 위 시계열 분석 방법으로 원유시장 가격을 예측하기에 모델이 선형적으로 단순화되어있고, 다양한 환경 변수들을 포함하지 않아 원유 가격의 변동성을 충분히 예측하지 못하는 치명적인 단점을 가진다.

최근 데이터 분석 기법의 정교화에 따라 인공지능 경향을 구성[6, 7] 하는 연구 또한 진행되었다. 위와 같은 방법으로 시계열 데이터 분석이 가지는 단순성과 선형성을 해결할 수 있다는 장점을 가지지만 충분한 데이터가 요구되고 모델의 복잡성으로 인해 모델의 직관적 구성 및 분석에 어려움이 있다.

이에 본 논문에서는 Gaussian Kernel Function Networks(GKFN)의 방법을 사용한 새로운 Time Series Analysis를 소개하여 단기, 장기적으로 일반화할 수 있는 모델을 구축하고자 한다. Gaussian Kernel은 Gaussian function을 커널로 사용하여 선형 모델에 활용하고자 하는 변수다. Gaussian Kernel Function Networks은 다음과 같이 표현된다.

$$\hat{f}(x) = \sum_{i=1}^m w_i \psi_i(x), \psi_i(x) = e^{-\|x - \mu\|^2 / 2\sigma_i^2} \quad (1)$$

Gaussian Function의 경우 기존 머신러닝 모델에서 활용되는 다른 함수들에 비해 Mapping Capacity가 높다[1]는 장점이 있다. 또한 Gaussian Function을 활용하여 시계열 데이터에 대한 Kernel의 적절한 개수를 추정하는 이론적 배경을 제시한 연구[2] 또한 적용하였다.

1.B. 연구의 기여

본 연구의 기여는 다음과 같다.

비선형성을 가진 시계열 분석 : 기존 선형성을 가진 시계열 모델로 가격을 예측한 연구와 달리, 원유시장의 복잡성을 대변해주는 비선형 모델을 구축하였다.

확률 이론에 기반한 모델 유연성 확대 : 원유 가격 시계열 모델에 Mapping Capacity가 높은 Gaussian Function을 적용함으로써 직관적인 모델의 수립 및 parameter 수정의 유연성을 높였다. 또한, Model Function 추정에 필요한 최적의 Kernel의 개수를 통계적 추정에 근거하여 추산하였다.

확장 기술 적용하여 정확도 향상 : 데이터 부족 문제를 해결하기 위해 추가적인 Data augmentation 방법으로 데이터를 확장하여 유연하게 적용시켰다. 또한, Rolling forecast 방법을 적용하여 test data의

추가 학습을 통해 모델 정확도를 향상시켰다.

모델의 확장성 검증 : 형성된 GKFN 모델에 일 데이터, 주 데이터, 월 데이터를 모두 적용하여 기간에 따른 모델의 정확도 손실 최소화를 검증하였다.

II. Gaussian Kernel Function Network의 구성

본 논문에서는 원유 가격에 대한 시계열 모델 근사를 위하여 GKFN를 이용한다. GKFN은 그림 1의 절차로 진행된다.[1,2,3]

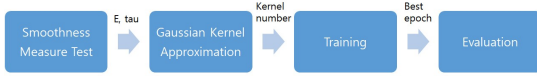


그림 1. Gaussian Kernel Function Networks 모델 구현 절차

Fig. 1. Modeling Process for Gaussian Kernel Function Networks

첫째, Smoothness Measure Test를 통해 학습에 용이한 하이퍼 파라미터들(τ , E)을 수학적으로 추산한다. 수학적으로 추산한 하이퍼 파라미터를 적용하여 Gaussian kernel regression model에 활용할 독립변수 데이터를 재구성한다.

둘째, 재구성한 독립변수 데이터를 활용하여 과적합을 방지하는 최적의 커널 개수를 확률적으로 추정한다. 이는 추정치의 noise variance를 통해 분석하여 최적의 커널 개수를 추정할 수 있다.

셋째, 최적의 커널 개수로 Gaussian kernel regression model을 구성 후 주어진 학습데이터를 활용하여 해당 모델의 가중치를 학습하는 과정이다. 이 과정에서 데이터 수가 부족한 문제를 해결하기 위해 Data augmentation 방법을 적용하여 데이터 수를 확장시켰다. Epoch 수는 100번까지 진행한 후 가장 높은 정확도를 가진 모델 가중치에 해당하는 Epoch로 결정하였다.

넷째, 학습한 모델을 Test data로 평가하는 과정이다. 이 과정에서 모델 정확도 향상을 위해 평가를 완료한 Test data를 Training data에 포함시켜 다시 모델의 가중치를 학습하여 업데이트하는 Rolling

Forecast 기법을 적용하였다.

2.A. Smoothness Measure Test

Gaussian Kernel Function Networks를 구성하기 위하여 첫째로 Smoothness Measure Test를 실행한다. Smoothness Measure Test는 모델을 원활한 학습이 가능한 함수의 형태로 구성하는 필요한 하이퍼 파라미터를 선정하는 과정이다. 위 과정을 통해 적절한 Embedding Dimension과 Interval을 정한다[8,9]. E 는 Embedding Dimension, 즉 다음 원유 가격을 예측하기 위해 독립변수로 활용할 이전 가격 데이터의 개수를 의미한다. τ 는 Interval, 즉 독립변수 간 시간 간격을 의미한다. GKFN 모델을 구성하는 E 차원의 독립변수 벡터를 $\vec{x}_{\tau,E}(k)$ 로 표현하는 것으로 가정하면 다음과 같다.

$$\vec{x}_{\tau,E}(k) = [x(k), x(k-\tau), \dots, x(k-(E-1)\tau)]^T \quad (2)$$

(2)는 현재 시점을 기준으로 τ 간격으로 E 개의 데이터를 원소로 하는 벡터이다. 이 벡터를 바탕으로 $k+P$ 기의 값이 함수 f 로 표현하는 것으로 가정하면, 다음과 같은 식으로 표현할 수 있다.

$$x(k+P) = f(\vec{x}_{\tau,E}(k)) \quad (3)$$

(3) 식에 근거하여, 데이터로부터 여러 쌍의 데이터 간 간격(τ)과 데이터의 개수(E)의 조합을 시도하는 것이 가능하다. 그 조합들 중 최적의 하이퍼 파라미터를 결정할 수 있다.

최적의 하이퍼 파라미터를 결정하는 요소는 구성된 함수 f 의 gradient다. 정의된 함수 f 의 각 점에서의 gradient는 식 (4)로 정의할 수 있다.

$$\Delta f(\vec{x}_{\tau,E}^r) = \frac{|f(\vec{x}_{\tau,E}^r(k)) - f(\vec{x}_{\tau,E}^r(k))|}{||\vec{x}_{\tau,E}^r(k) - \vec{x}_{\tau,E}^r(k)||} \quad (4)$$

(4) 식에서 분모는 두 벡터 사이의 Euclidean Distance를 표현하며, 분자의 경우 각각의 벡터에 대응되는 함수 값의 차이를 의미한다. r 의 경우 두 벡터가 얼마만큼 떨어져 있는 시계열 데이터인가를

의미한다. 본 논문에서는 $r=1$ 로 두고 smoothness measure를 측정한다.

(4) 식을 바탕으로 Smoothness Measure는 다음과 같이 정의된다.

$$S(\tau, E) = 1 - \frac{\sum_{k=(E-1)\tau}^{n-1} \Delta f(\vec{x}_{\tau, E}(k))}{n - (E-1)\tau} \quad (5)$$

(5) 식은 특정 하이퍼 파라미터로 구성된 모델에서 얼마만큼 gradient 값들의 평균이 작게 가지는지 보여주는 식이다. 이 식을 통해 하이퍼 파라미터(τ, E) Smoothness Measure(S)가 양수가 되는 가장 최초의 값을 선정하게 된다. 목표함수의 Gradient 평균이 1보다 낮아 학습을 용이하게 하는 (τ, E)를 Grid Search해 나가는 과정이다.

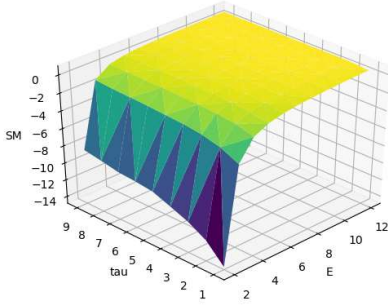


그림 2. (τ, E) 값을 결정하기 위한 Smoothness Test
Fig. 2. Smoothness measure test to estimate (τ, E)

2.B. 과적합 방지를 위한 noise variance 분석

Smoothness test 과정을 통해 주어진 (τ, E)의 값을 가장 잘 근사하는 GKFN을 구성하는 것이 다음의 목표이다. 추정된 식 $\hat{f}(x)$ 은 (1) 식으로 정의할 수 있다.

$$\hat{f}(x) = \sum_{i=1}^m w_i \psi_i(x), \quad \psi_i(x) = e^{-\|x - \mu_i\|^2 / 2\sigma_i^2} \quad (1)$$

(1) 식에서 m 은 커널의 개수, μ_i 과 σ_i 는 각각 커널의 평균과 분산을 의미한다.

그리고 과대적합을 방지하기 위하여 noise variance를 추정할 수 있다. 우선, 관측치와 예측치

간의 차이를 다음과 같은 식으로 표현 할 수 있다.

$$E[(x - \hat{x})^2] = E[(f - \hat{f})^2] + Var(\epsilon) \quad (6)$$

(6) 식에서의 오차는 커널 변수의 학습을 통해 줄여 나갈 수 있는 오차인 $E[(f - \hat{f})^2]$ 와 True model인 함수 f 에서도 관측하지 못하는 noise variance인 $Var(\epsilon)$ 로 구성되어 있다. 이 경우 noise variance 미만으로 오차를 줄이는 parameter 설정을 했을 때 과적합이 일어날 수 있다.

따라서 noise variance가 정규분포를 따른다고 가정한 뒤, noise의 추정값에 대한 confidence interval을 계산 할 수 있다. 우선 시계열 데이터간의 차를 분석하여 noise variance를 다음과 같이 식으로 표현할 수 있다.

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{k=1}^{n-1} (x(t_k) - x(t_{k-1}))^2 \quad (7)$$

noise가 정규분포를 따른다고 가정하였으므로, noise의 추정값은 chi-square distribution을 따른다. 이에 근거하여 추정 모델의 과적합 방지를 위하여 모분산을 추정할 수 있으며[3] 범위는 다음과 같다.

$$\frac{(n-1)\hat{\sigma}^2}{X_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)\hat{\sigma}^2}{X_{1-\alpha/2, n-1}^2} \quad (8)$$

따라서 주어진 모델에서 Training Error가 상기 식의 범위 내로 들어왔을 경우, 과적합이 시작되는 지점으로 추론 가능하다.

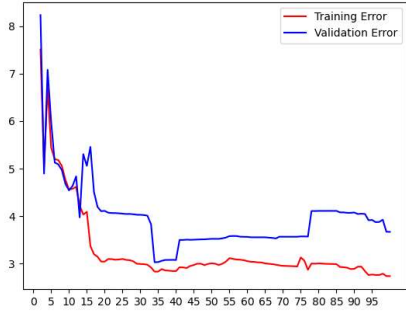


그림 3. 커널 개수에 따른 트레이닝 그래프
Fig. 3. Training graph according to kernel numbers

2.C. 모델 학습

2.C.1 모델 학습 과정

가우시안 커널 네트워크 모델의 학습은 총 세 단계로 구성된다[2]. 첫째로 noise variance estimate를 통해 최적의 커널 개수 m 을 지정한다(phase 1). Training error(Root Mean Square Error)가 noise variance estimate 범위 안에 들어오는 시점까지 커널의 개수를 늘려주게 된다. 이 과정은 섹션 2.B에 상세히 기술하였다.

둘째로, 커널 함수의 shape parameter를 조절하게 된다(phase 2). 가우시안 커널의 shape parameter는 function approximation의 basis가 되는 mean과 variance가 있다. 가우시안 커널의 mean은 함수의 구간을, variance는 함수의 곡률을 조정하며 주어진 데이터에 적합한 함수로 근사하게 된다.

마지막으로, 가우시안 커널 네트워크의 weight parameter를 조정하는 단계가 있다(phase 3). 이는 두 번째 단계에서 shape parameter가 조정되었기 때문에 기존의 weight를 다시 설정해 줄 필요가 있기에 지정되는 단계이다.

두 번째와 세 번째 단계는 모델의 주 학습과정으로 epoch를 100번으로 하여 반복적으로 weight를 조정해나갔다. epoch를 100번이 될 때까지 매 epoch마다 모델을 업데이트 시키고 나면, test set을 활용하여 초기 예측을 하여 정확도를 예측한다. 각 epoch에서 업데이트된 모델들 중 가장 정확도가 높은 epoch에서의 모델을 최종 모델로 선택하고 최종 평가 단계에 돌입한다.

2.C.2. Data Augmentation

Gaussian Kernel Function Networks Network은 Embedding Dimension(E)과 Kernel number(m)에 따라 많은 수의 모델의 가중치를 학습해야 하는 가능성이 생긴다. 그러나, 주 데이터, 월 데이터의 경우 높은 차원의 모델 가중치를 잘 학습시키기에 데이터 수가 충분하지 않는 문제가 발생된다. 이를 해결하기 위해 Data Augmentation이라는 방법을 적용하여 데이터를 확대 시켰다.

주간 분석(weekly analysis)의 경우, 일 데이터를 5일 간격으로 주 데이터처럼 구성한다. 예를 들어, 월요일~월요일로 한 데이터셋, 화요일~화요일로 한 데이터셋 등 5개의 데이터셋으로 재구성할 수 있다. 이와 같이 진행하는 경우 5배 이상의 데이터를 주간 분석에서 활용할 수 있다.

월간 분석(monthly analysis)의 경우, 일 데이터를 20일 간격으로 월 데이터처럼 구성한다. 예를 들어, 매월 1일로 이루어진 데이터셋, 매월 2일로 이루어진 데이터셋 등으로 재구성할 수 있다. 이는 주간 분석과 유사하게 월 데이터보다 20배 이상의 데이터를 월간 분석에서도 활용할 수 있어 더 높은 정확도를 기대할 수 있다.

2.C.3. P의 조정($P=4$)

월간 분석의 경우, Data Augmentation 외 데이터를 확보하는 다른 방법으로, P 를 조정하여 주 데이터를 활용하는 방법 또한 시도하였다. (2) 식과 (3) 식을 종합하여 나타내면 다음과 같다.

$$x(k+P) = f(x(k), x(k-\tau), \dots, x(k-(E-1)\tau)) \quad (8)$$

일반적인 모델 구현은 $P=1$ 로 설정하여 그 다음의 데이터를 예측하는 모델로 구현하는 것이다. 월간 분석을 위해 주 데이터를 활용하여 (8) 식에서 $P=4$ 로 설정하게 되면, 해당 모델은 4주 뒤, 즉 한 달 뒤의 원유가격을 예측하는 모델로 해석할 수 있다. 이와 같이 진행하는 경우, 주 데이터 개수만큼 학습데이터로 활용할 수 있기 때문에 월 데이터를

1) 원유 가격의 일 데이터는 주말을 제외한 월~금까지의 데이터로 구성된다. 1주일 간격은 월~월로 5일로 계산된다.

활용한 월간 분석보다 정확한 관측이 예상된다.

2.D. 모델 평가

본 논문에서는 일, 주, 월 단위의 예측을 시도하였다. 학습된 모델 가중치를 바탕으로 사전에 구성한 test data로 예측하였다.

평가 방법으로 활용한 지표는 RMSE, R-square, MAE이다.

$$R^2 = 1 - \frac{\sum_k^N (x(k) - f(\vec{x}_{\tau,E}(k)))^2}{\sum_k^n (x(k) - \bar{x}(k))^2}$$

$$RMSE = \sqrt{\frac{\sum_k^N (x(k) - f(\vec{x}_{\tau,E}(k)))^2}{N}}$$

$$MAE = \frac{\sum_k^N |x(k) - f(\vec{x}_{\tau,E}(k))|}{N} \quad (9)$$

평가를 위한 baseline model로 시계열 분석의 대표적인 모델인 ARIMA 모델을 사용하였다.

2.D.1. Rolling Forecast

최종 모델의 정확도를 더 향상시키기 위한 목적으로 test data를 적용하여 평가를 완료하면, test data를 training data로 이동하여 해당 데이터를 활용하여 학습을 시도하는 방법인 Rolling Forecast 기법을 적용하였다. 이는 ARIMA 모델에서 흔히 사용되는 기법으로, Test data를 모델 평가를 위한 데이터셋으로만 활용하지 않고, 평가 완료 후 training data로 활용하여 모델의 정확도를 높이기 위한 방법이다.

III. 원유 가격 예측 결과

본 논문의 데이터는 U.S. Energy Information Administration(EIA)에서 제공하고 있는 WTI 유가지수를 사용하였다. 이를 다음과 같이 일, 주, 월 단위로 데이터를 구성하였다.

표 1. 실험 데이터 세트

Table 1. Experimental data sets

데이터 종류	데이터 개수	데이터 범위
일	8,737	1986.01.02. - 2020.08.31.
주	1,809	1986.01.03. - 2020.08.28.
월	416	1986.01.01. - 2020.08.01.

일간, 주간, 월간 분석 시 Data Augmentation 등의 방법을 통해 표1의 데이터를 재구성하여 활용하기 위해서 동일한 데이터 기간 범위로 데이터셋을 구성하였다. 해당 데이터로 섹션 2에서의 모델 구현 단계인 Smoothness Measure Test, Kernel Approximation, Training, Evaluation(Prediction)의 단계를 거쳤다. Training 과정에서 데이터는 일, 주, 월 데이터 모두 80%를 사용하여 학습하였고, 나머지 20%의 데이터를 test data로 활용하여 평가 후 rolling forecast 방법을 적용하였다.

일간 분석의 경우, 일 데이터를 활용하여 GKFN 모델과 ARIMA 모델을 각각 구현하여 비교하였다. 주간 분석의 경우, 주 데이터를 활용한 GKFN 모델과 ARIMA 모델 구현 뿐만 아니라, 일 데이터를 활용한 Augmented GKFN 모델도 구현하여 비교하였다. 월간 분석의 경우, 월 데이터 수 부족 문제를 해결하기 위해 P=4와 Data Augmentation을 각각 시도하여 성능 차이를 비교하였다.



그림 4. 유가 예측 결과 그래프

Fig. 4. Plotted graph for oil price prediction

표 2. 일간 분석

Table 2. Daily Analysis

Model	RMSE	R2	MAE
GKFN	2.084	0.989	1.138
ARIMA	2.159	0.989	0.999

표 3. 주간 분석

Table 3. Weekly Analysis

Model	RMSE	R2	MAE
GKFN (weekly)	2.777	0.981	2.045
GKFN: Augmented (daily)	2.306	0.987	1.217
ARIMA (weekly)	2.454	0.985	1.691

*괄호 안은 분석을 위해 활용한 데이터셋의 종류를 의미한다.

표 4. 월간 분석

Table 4. Monthly Analysis

Model	RMSE	R2	MAE
GKFN (monthly)	5.812	0.914	4.580
GKFN: P=4 (weekly)	3.460	0.970	2.535
GKFN: Augmented (daily)	2.482	0.985	1.317
ARIMA (monthly)	5.397	0.926	4.202

*괄호 안은 분석을 위해 활용한 데이터셋의 종류를 의미한다.

일간 분석은 표2의 결과를 바탕으로 다음과 같이 분석할 수 있다. GKFN 모델과 ARIMA 모델과 동일한 R-square 결과가 나타났다. 하지만 RMSE는 GKFN 모델에서, MAE는 ARIMA 모델에서 우위를 가지는 것으로 나타났다. 이렇게 지표에 따른 모델 간 성능 차이는 해당 모델이 특정 시기의 data의 기 현상을 얼마만큼 잘 잡았는지에 따라 달라지는 현상이라 볼 수 있다. 하지만, 일간 분석의 경우 이마저도 차이가 거의 없어 GKFN과 ARIMA 모델이 유사한 정확도를 보여준다고 해석할 수 있다.

주간 분석은 표3의 결과를 바탕으로 다음과 같이 분석할 수 있다. 일반 GKFN 모델과 ARIMA 모델은 R-square가 98.1~98.5%로 나타난 것에 비해, Data augmentation을 통해 추가 확보한 데이터로 학습한 Augmented GKFN의 경우 98.7%의 R-square 결과가 나왔고, 다른 지표도 기존 모델보다 좋은 성능을 보였다. Augmented GKFN의 결과가 일간 분석의 결과

와 차이가 없다는 결과도 확인할 수 있다.

월간 분석은 표4의 결과를 바탕으로 다음과 같이 분석할 수 있다. 일반 GKFN 모델과 ARIMA 모델은 R-square가 91%~92%로 나타난 것에 비해, 파라미터를 일부 변형(P=4)한 GKFN 모델은 97%, Augmented GKFN 모델은 98.5%의 R-square 결과가 나왔다. 그 외 다른 지표도 데이터 수를 확대한 변형 모델들이 기존 모델보다 좋은 성능을 보임을 확인할 수 있다. 특히, Augmented GKFN 모델의 결과를 일간 분석의 결과와 비교하였을 때 차이가 없는 것으로 보았을 때, Augmented GKFN이 단기, 장기와 무관하게 모델 정확도의 손실이 거의 없는 것을 알 수 있다.

IV. 결 론

본 논문에서는 원유가격 예측을 위한 모델로 비선형 모델인 Gaussian Kernel Function Networks 방법으로 시계열 분석을 시도하였다. GKFN을 통해 보다 직관적으로 E, τ , Kernel Number와 같은 Hyperparameter들을 설정 할 수 있었으며, 모분산의 통계적 추산에 근거하여 과적합 방지 기술을 적용하였다. 이를 통해 모델의 복잡도를 줄이고 직관성과 일반화 가능성을 높일 수 있었다. 이에 더하여 Rolling Forecast, Data Augmentation과 같은 다양한 확장 기술이 적용 가능한 것을 보여주며 유가 분석에 있어 GKFN 모델 구성이 가지는 유연성이라는 이점을 극대화하였다. GKFN 모델을 통해 예측한 결과로 일간 분석, 주간 분석, 월간 분석을 시도하였고, baseline method인 ARIMA 모델과 다르게 단기에서 장기로 모델을 구성하여도 모델 정확도의 손실이 일어나지 않았다. 이렇게 각기 다른 성질의 데이터에도 높은 정확도와 확장성을 가진 시계열 분석 모델을 구축할 수 있었다.

참 고 문 헌

- [1] Sukhan Lee and Rhee M. Kil, “A Gaussian Potential Function Network With Hierarchically Self-Organizing Learning”, Neural Networks, Vol. 4, pp. 207-224, 1991.
- [2] Rhee M. Kil, “Function Approximation Based on a Network with Kernel Functions of Bounds and Locality: an Approach of Non-Parametric Estimation”, ETRI Journal, Vol. 15, No 2, October 1993.
- [3] Dong Kyu Kim and Rhee M. Kil, “Stock Price Prediction Based on a Network with Gaussian Kernel Functions”, ICONIP 2013, Part II, LNCS 8227, pp. 705-712, 2013.
- [4] C. Baumeister, L. Kilian, “Real-time forecasts of the real price of oil”, Journal of Business & Economics Statistics, 30, 326-336, 2012.
- [5] Angela W. W. He, Jerry T. K. Kwok, Alan T.K. Wan, “An empirical model of daily highs and lows of West Texas Intermediate crude oil prices”, Energy Economics, Elsevier, vol. 32(6), pages 1499-1506, Nov, 2010.
- [6] Ganapathy S. Natarajan, Aishwarya Ashok, “Multivariate Forecasting of Crude oil Spot Prices using Neural Networks”, arXiv:1811.08963 [cs.LG], Nov, 2018.
- [7] Indranil SenGupta, William Nganje, Erik Hanson, “Refinements of Brandorff-Nielson and Shepard model: an analysis of crude oil price with machine learning”, arXiv:1911.13300 [q-fin.ST], Mar, 2020.
- [8] Packard, N.H., Crutchfield, J.P., Farmer, J.D., Shaw, R.S., “Geometry from a TimeSeries. Phys.”, Rev. Lett. 45, pp. 712-716, 1980.
- [9] Takens, F.: “Detecting strange attractors in turbulence. In: Rand, D.A., Young, L.S. (eds.)”, EAMT-WS 1993. Lecture Notes in Mathematics, vol. 898, pp. 366-381. Springer-Verlag, Berlin 1981.