Journal of the Korean Data & Information Science Society
2019, **30**(3), 645–654
http://dx.doi.org/10.7465/jkdi.2019.30.3.645
한국데이터정보과학회지

디노이징 필터와 LSTM을 활용한 KOSPI200 선물지수 예측

이낙영[1] · 오경주[2]
[12]연세대학교 산업공학과
접수 2019년 4월 8일, 수정 2019년 5월 8일, 게재확정 2019년 5월 15일

요 약

딥러닝 모형을 통해 금융시장을 예측하는 연구는 활발하지만 디노이징 필터를 적용하여 금융 데이터의 노이즈를 제거함으로써 예측 모형의 성능을 높이는 연구는 거의 이루어지지 않고 있다. 따라서, 본 연구의 목적은 디노이징 필터를 사용하여 데이터의 노이즈를 제거한 후 시계열 예측에 유용한 딥러닝 모형인 LSTM의 예측 성능을 높이는 것이다. KOSPI200 선물지수의 일봉과 30분봉 데이터를 이용해 실증분석을 하였다. 디노이징 필터를 적용한 예측 모형의 성능이 기존 LSTM보다 전체 기간 실험과 슬라이딩 윈도우 실험을 통해 우수함을 입증하였다. 또한, 제안한 디노이징 필터 중 사비츠키-골레이 필터가 이동평균 필터보다 예측 모형 성능 향상에 유용함을 확인하였다. 향후, 디노이징 필터가 다양한 딥러닝 모형의 예측 성능 향상에 사용될 수 있음을 기대한다.

주요용어: 디노이징 필터, 슬라이딩 윈도우, KOSPI200 선물지수, LSTM.

1. 서론

KOSPI200 선물 (Futures)은 KOSPI200 주가지수를 기초자산으로 하는 파생상품이며, 증거금 제도로 인해 레버리지 (leverage)투자가 가능하다. KOSPI200 선물의 일평균 계약수는 2018년 26만 계약으로 전년도에 비해 약 32% 증가하였고 일평균 계약금액은 2018년 20조원으로 전년도에 비해 약 14% 증가하였다 (한국거래소). 이를 통해, 많은 투자자들이 KOSPI200 선물거래에 관심을 가지고 있다는 것을 알 수 있다. 또한, 선물지수를 예측하려는 연구도 활발하다. Lee (2014)는 다양한 확률모형을 사용하여 KOSPI200 선물지수 예측에 관한 연구를 진행하였고 Kim 등 (2015)은 머신러닝 알고리즘인 K-NN의 KOSPI200 선물지수 예측 성능을 확인하였다. 하지만, 금융데이터는 변동성이 크고 노이즈가 많기 때문에 예측이 어렵다. 이러한 문제를 해결하기 위해 최근에는 다양한 딥러닝 (deep learning) 모형을 통해 금융시장을 예측하는 연구가 활발하다. Lee (2017)는 딥러닝과 기술적 지표를 통해 KOSPI200 주가지수의 방향성 예측을 시도하였고 Shin 등 (2017)은 LSTM이 기존 심층신경망에 비해 주가 예측 성능이 더 좋다는 것을 보였다. 하지만, 이러한 연구들에도 불구하고 디노이징 필터를 통해 데이터의 노이즈를 제거한 후 예측 모형의 성능을 높이는 연구는 매우 미흡한 실정이다.

본 연구에서는 신호처리에 사용되는 디노이징 필터를 통해 데이터의 노이즈를 제거한 후, 시계열 데이터 예측에 효과적인 LSTM을 활용하여 KOSPI200 선물지수 일봉과 30분봉을 예측하였고 기존 LSTM의 예측 성능보다 향상됨을 확인하였다. 향후, 다양한 딥러닝 모형의 예측 성능 향상에 디노이징 필터가 사용될 수 있음을 기대한다.

본 논문의 구성은 다음과 같다. 2절에서는 디노이징 필터와 LSTM에 대해 설명하였다. 3절에서는 실증분석에 사용된 데이터를 소개하고 기존 LSTM 예측 모형과 디노이징 필터를 적용한 예측 모형의 성능을 비교하였다. 마지막으로, 4절에서는 실증분석 결과 및 기대효과에 대해 서술하였다.

[1] (03722) 서울특별시 서대문구 연세로 50, 연세대학교 산업공학과, 석사과정.
[2] 교신저자: (03722) 서울특별시 서대문구 연세로 50, 연세대학교 산업공학과, 교수.
E-mail: johanoh@yonsei.ac.kr

2. 연구 배경

2.1. 이동평균 필터

이동평균 필터는 신호처리에서 노이즈 제거에 사용되는 가장 기본적인 필터이다. 이동평균은 투자전략을 세울 때 가격의 추세를 확인하기 위한 방법으로 많이 사용되는 기술적 지표이기도 하다. 이동평균 필터는 데이터의 특정 개수를 일정한 간격만큼 이동시켜가면서 계산한 평균이며, 수식은 아래 (2.1)과 같다.

$$\bar{x_k} = \frac{x_{k-n+1} + x_{k-n+2} + \cdots + x_k}{n}. \tag{2.1}$$

이동평균 필터는 지속적으로 반복되는 노이즈 제거와 급작스런 변화에 유용하지만 데이터의 진폭이 큰 경우엔 각 데이터의 가중치가 동일하여 성능이 제한적이다 (Jung과 Kim, 2014).

2.2. 사비츠키-골레이 필터

사비츠키-골레이 (savitzky-golay)필터는 특정 차수의 다항식을 최소제곱법 (method of least squares)를 사용하여 데이터에 피팅 (fitting)함으로써 데이터의 손상을 최소화하며 노이즈를 제거하는 방법이다 (Savitzky와 Golay, 1964).이동평균 필터에 비해 데이터의 진폭의 폭과 너비를 더 잘 보존하며, 신호처리와 이미지 처리 분야에서 사용되고 있는 필터이다 (Azami 등, 2012; Hargittai, 2005; Jalab과 Ibrahim, 2013).

노이즈가 제거된 데이터 $p(n)$은 최소제곱법을 통해 구해진 계수 a_k를 통해 다음과 같이 표현된다 (Schafer, 2011).

$$p(n) = \sum_{k=0}^{N} a_k n^k.$$

N은 N차 다항식을, n은 데이터 포인트 (point)를 의미한다. $n = 2M + 1$개이며, M은 데이터 포인트의 중앙값을 의미한다. 최소제곱법을 통해 계수 a_k를 구하는 식은 (2.2)와 같고 $x[n]$은 데이터 포인트의 값을 의미한다.

$$Minimize \sum_{n=-M}^{M} (p(n) - x[n])^2. \tag{2.2}$$

본 연구에서는 2차 다항식을 사용하여 실증분석을 진행하였다.

2.3. LSTM

LSTM (Long short term memory)는 순환신경망의 기울기 소실 (vanishing gradient)문제와 느린 학습 속도를 극복하기 위해 제안된 알고리즘으로, 순환신경망의 은닉층에 LSTM 셀을 추가한 것이다 (Hochreiter와 Schmidhuber, 1997). Figure 2.1은 LSTM의 구조를 나타낸 그림이다. LSTM 셀은 입력 게이트 (input gate; i_t), 망각 게이트 (forget gate; f_t), 출력 게이트 (output gate; o_t)와 단기 상태 (short term state; s_t), 장기 상태 (long term state; l_t)로 이루어져 있다 (Graves와 Schmidhuber, 2005). LSTM의 구조를 간단히 설명하면, 입력 정보 (x_t)를 입력받으면 i_t를 통해 어떤 정보를 받아

들일지 정하고 f_t에서 필요없는 정보를 잊어버린 후 l_{t-1}의 정보와 계산된다. s_{t-1}의 정보와 o_t를 통해 계산된 정보와 이전 정보가 계산되어 출력 정보 (y_t)를 내보낸다. 딥러닝 모형에서 학습을 빠르고 안정적이게 하기 위한 최적화 기법으로 Adagrad, RMSProp, Adam 등이 있다. Adam 최적화기는 Adagrad의 기울기 (gradient)를 잘 찾는 이점과 RMSProp의 비정상적 목적함수를 다루는 데 용이한 이점을 결합한 최적화 기법이다 (Kingma와 Ba, 2014). 본 연구에서는 3개의 LSTM 셀과 Adam 최적화기를 사용하여 실증분석을 진행하였다.

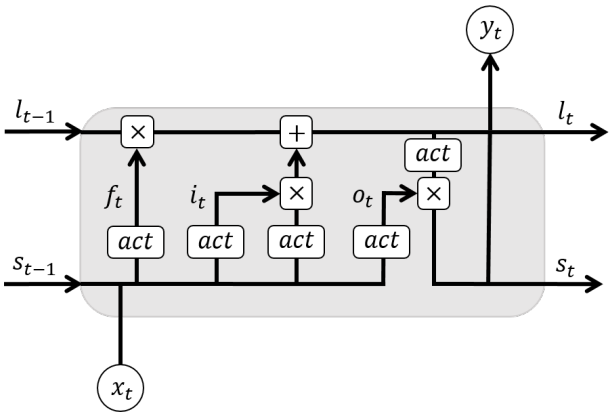


Figure 2.1 LSTM structure

3. 실증 분석

3.1. 데이터

본 연구에서는 일봉 실험과 30분봉 실험에 2007년 1월 2일부터 2018년 12월 28일까지의 데이터와 2017년 1월 2일부터 2018년 12월 28일까지의 KOSPI200 선물지수 시가, 고가, 저가, 종가, 거래량을 사용하였다. 실증분석에 사용한 데이터는 디노이징 필터를 통해 노이즈를 제거한 후 LSTM의 입력변수로 사용되며 최종적으로 종가 지수를 예측한다. 일반적으로, 디노이징 필터는 스무딩 (smoothing)을 통해 데이터를 매끄럽게 함으로써 데이터의 노이즈를 제거한다. Figure 3.1을 통해 필터된 데이터가 기존 원 데이터에 비해 스무딩된 것을 확인할 수 있다.

Figure 3.2 (a)와 (b)는 각각 원 데이터와 필터된 데이터의 차이와 그 차이의 히스토그램을 나타낸 그림이다. KOSPI200 선물지수 데이터와 사비츠키-골레이 필터를 적용한 데이터의 차이는 평균 0.0012, 표준편차 2.25의 값을 가진다. 이를 통해 잔차의 히스토그램이 정규분포의 모양을 가지는 노이즈를 제거했다는 것을 확인할 수 있다.

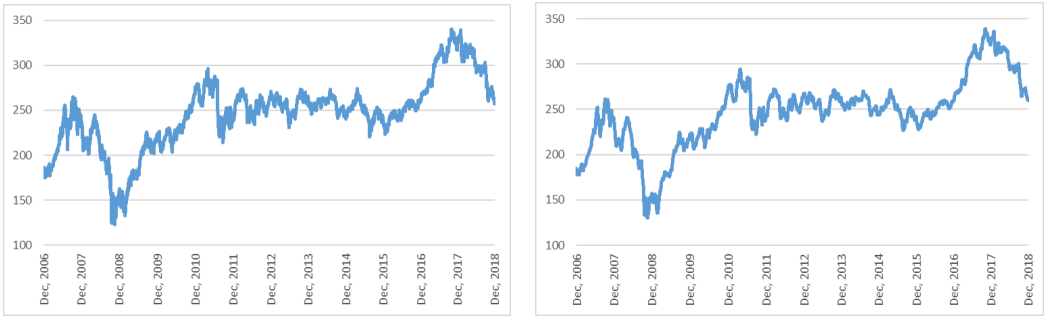
금융시계열에서 단기(5일-1주), 중기(20일-1달), 장기(60일-1분기) 이동평균 값을 사용하는데 윈도우 값이 크면 데이터 래깅 (lagging) 현상이 발생하여 이동평균 필터의 윈도우 값은 5로 설정하였으며, 사비츠키-골레이 필터의 윈도우 값이 크면 데이터 손상이 발생하고 작으면 노이즈 제거 성능이 떨어지기 때문에 15로 설정하였다 (Chen 등, 2004). 상대적으로 변동성이 작은 30분봉 데이터의 경우, 일봉 데이터보다 낮은 윈도우 수를 적용하였다.

본 연구의 실증 분석은 Python의 Tensorflow를 사용하여 진행하였다. 일반적으로, 딥러닝 모형에서

노드의 개수가 적으면 학습이 잘 이루어지지 않고 많으면 과적합 (overfitting)이 발생하기 때문에 적절한 노드의 개수를 정하기 위해 노드의 개수를 변화시키면서 실험을 진행하였다. 본 연구에서는 Table 3.1의 결과에 따라 가장 큰 예측 성능 향상을 보인 30개의 노드를 사용하여 실험을 진행하였다. LSTM의 학습률 (learning rate)은 값이 크거나 작으면 정확한 값을 찾지 못하거나 시간이 오래 걸리는 문제가 발생하여 0.05로 설정하였고 1500번의 학습을 수행하였다 (Huang 등, 2014).

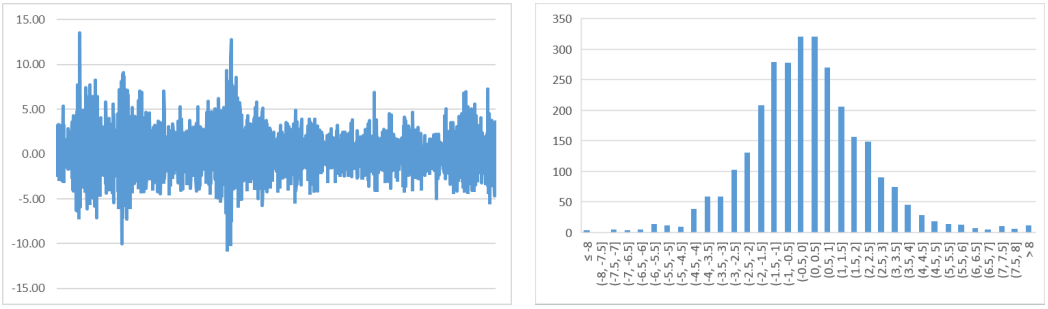
Table 3.1 Prediction performance by nodes (RMSE)

Node	LSTM	MA+LSTM	SV+LSTM
10	5.79	7.97	5.23
15	5.75	4.33	7.08
20	5.98	5.04	3.56
30	6.79	4.53	2.89
60	10.45	5.14	9.44
120	12.82	23.37	14.96



(a) Close price (b) Denoised close price

Figure 3.1 Daily close price of KOSPI200 futures



(a) Graph (b) Histogram

Figure 3.2 Difference between raw data and filtered data

3.2. 예측 모형 평가

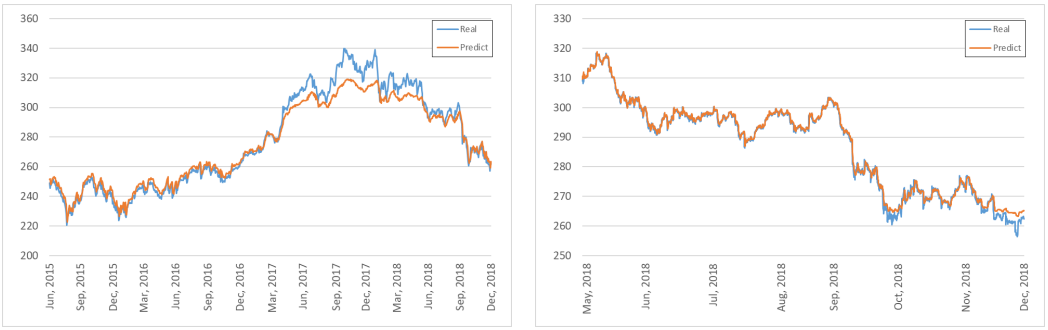
본 연구에서는 평균제곱근오차 (root mean square error; RMSE)와 평균절대비율오차 (mean absolute percentage error; MAPE)를 사용하여 예측 모형의 성과를 평가하였다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \quad MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{3.1}$$

식 (3.1)에서 y_i와 $\hat{y}_i$는 각각 실제값과 예측값을 나타내며, n은 test set의 개수를 의미한다. RMSE와 MAPE는 값이 작을수록, 예측 모형의 성능이 좋다고 할 수 있다. 디노이징 필터를 적용한 예측 모형의 성능 평가는 노이즈를 제거한 데이터를 통해 나온 예측값과 실제값을 가지고 이루어진다.

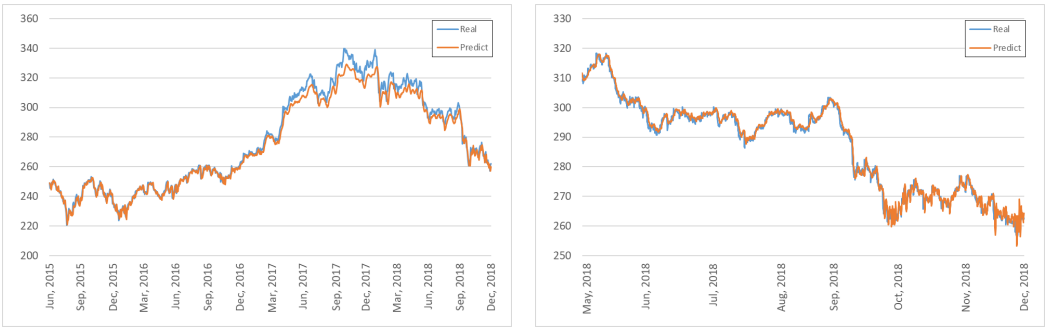
3.3. 전체 기간 실험

전체 기간 실험은 일봉, 30분봉 데이터를 이용하여 진행하였다. 2967개의 일봉 데이터와 6799개의 30분봉 데이터 중 70%가 training set, 30%가 test set으로 사용되었다. Figure 3.3~3.5는 LSTM, 이동평균 필터를 적용한 예측 모형, 사비츠키-골레이 필터를 적용한 예측 모형의 일봉과 30분봉 예측 그래프이다. Table 3.2는 전체 기간에 대한 예측성능을 나타낸다. 일봉의 경우, 이동평균 필터와 사비츠키-골레이 필터를 적용한 예측 모형의 성능이 LSTM보다 향상된 것을 확인할 수 있으며 30분봉의 경우, 사비츠키-골레이 필터를 적용한 예측 모형의 성능이 LSTM과 이동평균 필터를 적용한 예측 모형보다 향상된 것을 확인할 수 있다. 이를 통해, 변동성이 큰 일봉 데이터에서 이동평균 필터와 사비츠키-골레이 필터를 적용한 예측 모형의 성능 향상을 확인할 수 있었고 상대적으로 변동성이 작은 30분봉 데이터에서 사비츠키-골레이 필터를 적용한 예측 모형의 성능 향상을 확인할 수 있었다.



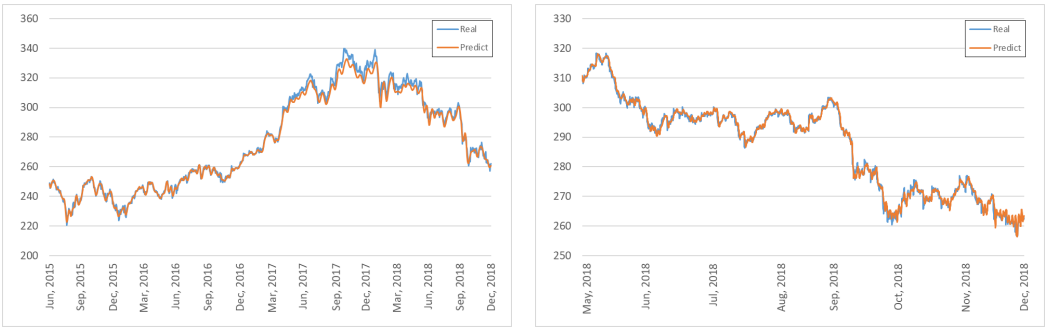
(a) Daily (b) 30 min

Figure 3.3 LSTM real-predict graph



(a) Daily (b) 30 min

Figure 3.4 MA+LSTM real-predict graph



(a) Daily (b) 30 min

Figure 3.5 SG+LSTM real-predict graph

Table 3.2 Prediction performance (left : daily, right : 30 min)

Model	RMSE	MAPE
LSTM	6.79	1.69
MA+LSTM	4.53	1.15
SG+LSTM	2.89	0.70

Model	RMSE	MAPE
LSTM	1.38	0.32
MA+LSTM	1.60	0.41
SG+LSTM	1.14	0.29

3.4. 슬라이딩 윈도우 실험

모델의 예측 성능 강건성 (robustness)를 위해 슬라이딩 윈도우 (sliding window) 실험을 진행하였다. 슬라이딩 윈도우란 training set과 test set을 일정 부분으로 나눈 후 test set만큼 이동하며 실험을 진행하는 것이다. 시계열 자료 분석 및 예측에 유용하며 데이터의 일부만 training에 사용함으로써, 과거데이터에 의한 영향이 줄어듦을 반영할 수 있다 (Jo 등, 2018; Kim과 Oh, 2019). 일봉과 30분봉의 슬라이딩 윈도우 실험은 각각 training set 5년 + test set 1년, training set 1년 + test set 1개월로 설정하여 진행하였다. Table 3.3과 Table 3.4는 기간별 test set과 전체 test set에 대한 일봉과 30분봉 예

측 성능을 나타낸다. 일봉과 30분봉 슬라이딩 윈도우 실험에서 이동평균 필터를 적용한 예측 모형의 성능이 LSTM에 비해 크게 향상되지 않았던 반면에 사비츠키-골레이 필터를 적용한 예측 모형의 성능은 LSTM과 이동평균 필터를 적용한 예측 모형에 비해 기간별 test set과 전체 test set에서 크게 향상된 것을 확인할 수 있었다.

Table 3.3 Daily prediction performance (top : RMSE, bottom : MAPE)

Model	2012	2013	2014	2015	2016	2017	2018	Total
LSTM	2.89	2.70	2.07	2.74	2.30	4.28	3.20	2.96
MA+LSTM	3.32	2.55	2.57	2.20	2.30	3.74	3.15	2.88
SG+LSTM	1.93	1.71	1.82	2.65	2.16	3.32	2.18	2.31

Model	2012	2013	2014	2015	2016	2017	2018	Total
LSTM	0.86	0.80	0.62	0.83	0.72	1.06	0.80	0.81
MA+LSTM	1.03	0.75	0.78	0.68	0.68	1.03	0.79	0.82
SG+LSTM	0.61	0.54	0.57	0.79	0.68	0.90	0.57	0.67

Table 3.4 30 min prediction performance (top : RMSE, bottom : MAPE)

Model	1	2	3	4	5	6	7	8	9	10	11	12	Total
LSTM	0.69	1.77	1.35	0.89	0.63	0.68	0.66	0.61	0.60	1.01	1.48	0.94	1.01
MA+LSTM	0.79	1.58	1.43	0.92	0.66	0.71	0.70	0.62	0.72	1.02	1.17	1.00	0.98
SG+LSTM	0.41	1.15	0.86	0.74	0.39	0.53	0.45	0.57	0.36	0.75	0.70	0.68	0.67

Model	1	2	3	4	5	6	7	8	9	10	11	12	Total
LSTM	0.15	0.46	0.27	0.18	0.13	0.16	0.16	0.13	0.13	0.24	0.32	0.23	0.21
MA+LSTM	0.17	0.39	0.29	0.19	0.14	0.17	0.17	0.14	0.16	0.25	0.28	0.25	0.21
SG+LSTM	0.10	0.30	0.21	0.17	0.09	0.14	0.12	0.15	0.09	0.21	0.21	0.18	0.16

Table 3.5와 Table 3.6은 사비츠키-골레이 필터를 적용한 예측 성능이 LSTM에 비해 유의한 차이가 있는지를 검증하기 위해 t-test를 실시한 결과다. 충분한 표본을통한 검정을 위해 일봉의 월별 예측 성능과 30분봉의 주별 예측 성능의 통계적 검정을 진행하였다. Table 3.5와 Table 3.6에서 알 수 있듯이 두 실험 모두 LSTM과 사비츠키-골레이 필터를 적용한 예측 모형의 RMSE와 MAPE의 평균이 통계적 유의수준 하에서 차이가 있는 것으로 나타났다. 이를 통해, 사비츠키-골레이 필터를 적용한 예측 모형의 성능이 기존 LSTM 모델에 비해 높다고 해석할 수 있다.

Table 3.5 Daily t-test result

	LSTM	SG+LSTM
Observations	84	84
RMSE	t=3.68***	
Mean (Variance)	2.73 (1.39)	2.12 (0.92)
MAPE	t=2.69***	
Mean (Variance)	0.81 (0.13)	0.69 (0.12)

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 3.6 30 min t-test result

	LSTM	SG+LSTM
Observations	52	52
RMSE	t=3.77***	
Mean (Variance)	0.90 (0.22)	0.62 (0.07)
MAPE	t=2.29**	
Mean (Variance)	0.21 (0.02)	0.16 (0.01)

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$

4. 결론

본 연구는 디노이징 필터를 통해 데이터의 노이즈를 제거함으로써, LSTM 예측 모형의 성능을 높이고자 하였다. 연구결과, 디노이징 필터를 적용한 예측 모형이 기존 LSTM보다 성능이 향상되었다. 또한, 일봉 데이터에서만 성능 향상을 보인 이동평균 필터와 달리 사비츠키-골레이 필터를 적용한 예측 모형이 30분봉 데이터에서도 통계적으로 유의미한 성능 향상을 보이는 것을 실증적으로 입증하였다. 본 연구는 디노이징 필터의 윈도우를 특정 값으로 고정한 채 딥러닝 모형을 적용한 것에 대한 한계가 있다. 또한, 본 연구에서 사용한 데이터 이외에 금리, 물가 상승률, 환율 등의 거시경제변수를 사용한다면 더 좋은 예측 성능을 얻을 것이라 기대된다. 향후 디노이징 필터는 LSTM 이외의 딥러닝 모형의 예측 성능을 높이는 데 활용 가능할 것이라 생각된다.

References

Azami, H., Mohammadi, K. and Bozorgtabar, B. (2012). An improved signal segmentation using moving average and savitzky-golay filter. *Journal of Signal and Information Processing*, **3**, 39-44.

Chen, J., Jonsson, P., Tamura, M., Gu, Z., Matsushita, B. and Eklundh, L. (2004). A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter. *Remote sensing of Environment*, **91**, 332-344.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, **18**, 602-610.

Hargittai, S. (2005). Savitzky-golay least-squares polynomial filters in ECG signal processing. *Computers in Cardiology*, **32**, 763-766.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**, 1735-1780.

Huang, G. B., Zhu, Q. Y. and Siew, C. K. (2004). Extreme learning machine: A new learning scheme of feedforward neural networks. *Neural Networks*, **2**, 985-990.

Jalab, H.A. and Ibrahim, R.W. (2013). Texture enhancement based on the savitzky-golay fractional differential operator. *Mathematical Problems in Engineering*, **2013**, 1-8.

Jo, K. H., Jeong, S. H., Kim, K. S. and Oh, K. J. (2018). Scoring model to determine trade timing based on genetic algorithm. *Journal of Korean Data & Information Science Society*, **29**, 735-745.

Jung, I. B. and Kim, K. H. (2014). A study of PPG wave and pulse measurement on radial artery using digital potentiometer and exponentially weighted moving average filter. *The Transactions of the Korean Institute of Electrical Engineers*, **63**, 962-967.

Kim, E. C. and Oh, K. J. (2019). Asset allocation strategy using hidden Markov model and genetic algorithm. *Journal of Korean Data & Information Science Society*, **30**, 33-44.

Kim, M. H., Lee, S. H. and Shin, D. H. (2015). Predictability test of k-nearest neighbors (K-NN) algorithm : Application to the KOSPI200 futures. *Korean Journal of Business Administration*, **28**, 2613-2633.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1-15.

Lee, H. S. (2014). Forecasting the prices of KOSPI200 index future. *Korean Journal of Business Administration*, **27**, 2165-2179.

Lee, W. S. (2017). A deep learning analysis of the KOSPI's directions. *Journal of Korean Data & Information Science Society*, **28**, 287-295.

Savitzky, A. and Golay, MJE. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, **36**, 1627-1639.

Shin, D. H., Choi, K. H. and Kim, C. B. (2017). Deep learning model for prediction rate improvement of stock price using RNN and LSTM. *Korean Institute of Information Technology*, **15**, 9-16.

Schafer, R. W. (2011). What is a savitzky-golay filter. *IEEE Signal Processing Magazine*, **28**, 111-117.

Journal of the Korean Data & Information Science Society
2019, **30**(3), 645–654

http://dx.doi.org/10.7465/jkdi.2019.30.3.645
한국데이터정보과학회지

KOSPI200 futures index prediction using denoising filter and LSTM

Nak Young Lee[1] · Kyong Joo Oh[2]

[12]Department of Industrial Engineering, Yonsei University

Received 8 April 2019, revised 8 May 2019, accepted 15 May 2019

Abstract

There has been many studies which predict the financial market using the deep learning model. However, there has been few studies which apply the denoising filter that improves the performance of predictions removing the noise of financial data. Therefore, the purpose of this study is to apply denoising filter to remove noise from data and then to improve the prediction performance of long short term memory, a deep learning model which is useful for time series prediction. We conducted an empirical analysis using daily and 30 min KOSPI200 futures index data. It is proven that the performance of prediction model using denoising filter is superior to that of the previous long short term memory for the whole period and the sliding window experiment. Also, we confirmed that savitzky-golay filter is more useful for improving the prediction model performance than moving average filter. In the future, denosing filter may be used to improve the prediction performance of various deep learning models.

Keywords: Denoising filter, KOSPI200 future index, LSTM, sliding window.

[1] Graduate student, Department of Industrial Engineering, Yonsei University, Seoul 03722, Korea.
[2] Corresponding author: Professor, Department of Industrial Engineering, Yonsei University, Seoul 03722, Korea. E-mail: johanoh@yonsei.ac.kr