공학석사 학위논문

LSTM과 양방향 순환신경망을 이용한 주가 예측모델 비교연구

A Comparative Study on Stock Price Forecasting Models Using LSTM and Bidirectional Neural Networks

2019년 2월

서울과학기술대학교 일반대학원

SW분석·설계학과

이 종 혁

표지측면(등표지)

LSTM과 양방향 순환신경망을 이용한 주가 예측모델 비교연구

A Comparative Study on Stock Price Forecasting Models Using LSTM and Bidirectional Neural Networks

지도교수 국광호

이 논문을 공학석사 학위논문으로 제출함

2019년 2월

서울과학기술대학교 일반대학원

SW분석·설계학과

이 종 혁

이종혁의 공학석사 학위논문을 인준함

2019년 2월

심사위원장 장성용 (인)

심사위원 국광호 (인)

심사위원 김우제 (인)

요 약

제 목 : LSTM과 양방향 순환신경망을 이용한 주가 예측모델 비교연구

시계열 데이터(Time-series Data)는 주식, 매출액, 실업률 등과 같이 시간적 순서를 가지는 데이터를 의미한다. 현재 직면한 문제를 해결하려는 모든 조직에게 시계열 데이터의 관찰, 분석 더 나아가 데이터 예측은 필수다. 데이터 예측을 통해 조직 내외 의사결정 문제를 해결할 수 있을 뿐만 아니라 의사결정의 결과로서의 계획, 투자, 성과를 향상시킬 수 있으며 투자 기회비용, 재고비용, 물류비용 등과 같은 비용의 절감을 할 수 있기 때문이다.

시계열 데이터 중 주식 데이터를 다루는 주식 시장은 불특정 다수가 존재하며, 다양한 변수가 작용하는 시장으로 예측하기 어려운 것이 특징이다. 주식 시장을 예측하기 위한 많은 노력이 있지만, 선형적 예측으로는 한계가 있다. 전통적인 수학 모델의 한계를 극복하기 위해 인공신경망을 활용한 기계학습이 제안되고 있다. 신경망 모형 중 순환신경망(Recurrent Neural Network)은 일반 신경망에 시계열 개념을 추가한 것으로, 은닉계층에 이전 정보를 기억시킬 수 있는 장점이 있다. 순환신경망을 통해 과거의 지수 가격이 미래의 지수 가격에 어떻게 영향을 미치는지를 예측한다. 본 논문은 시계열 데이터 중 주식 예측모델을 다룬다. 이에 관한 다양한 연구가 있으며 특히, 양방향 순환신경망을 이용한 주식 가격 예측 모형 연구가 있지만 가격과 거래량에 한정된 데이터를 사용한 모델링을 했다. 본 논문에서는 가격과 거래량뿐만 아니라 가격 보조지표, 거래량 보조지표, 추세 보조지표, 채널 보조지표를 추가해 모델 간의 비교를 할 것이다. 각 모델 간의 비교를 통해 어떤 보조지표를 사용한 모델이 가장 좋은 성능을 내는지 평가할 수 있으며 이를 통해 모델링할 때, 가장 좋은 지표 조합을 알 수 있을 것이다. 본 실험을 통해 얻은 모델이 예측한 값과 예측하기 전날의 종가와 비교를 통해 상승/하락 예측의 정확도를 검증하였다. 또한 주식 가격의 상승/하락 변동이 크게 예측되는 경우에는 상승/하락 예측의 정확도가 더 높아질 수 있을 것이라는 가정하에 상승/하락률이 각각 0.5%, 1%, 2% 이상이 되는 경우만을 고려하여 예측의 정확도를 검증하였다. 이를 통해 본 논문에서 제안하는 모델이 주식의 상승/하락을 비교적 정확하게 예측할 수 있음을 보였다.

목 차

요약 ····· i
표목차 ····· iii
그림목차 ····· iii

I. 서 론 ····· 1
1. 연구의 배경 및 목적 ····· 1
2. 논문의 구성 ····· 2

Ⅱ. 관련 연구 ····· 3
1. 기술적 분석 ····· 3
2. 인공신경망 ····· 5
3. 순환신경망(Recurrent Neural Networks) ····· 7
4. Long Short-Term Memory ····· 9
5. 양방향 순환신경망(Bidirectional Recurrent Neural Networks) ····· 10

Ⅲ. 실험 및 성능평가 ····· 12
1. 실험 환경 ····· 12
2. 성능 평가 ····· 14
3. 성능 검증 ····· 22

Ⅳ. 결 론 ····· 24

참고문헌 ·····
영문초록(Abstract) ·····
감사의 글

표 목 차

Table 2.1 가격 보조지표 ····· 3
Table 2.2 거래량 보조지표 ····· 3
Table 2.3 추세 보조지표 ····· 4
Table 2.4 채널 보조지표 ····· 4
Table 3.1 양방향 순환신경망 계층 비교 ····· 13
Table 3.2 실험환경 ····· 13
Table 3.3 Case 분류 ····· 14
Table 3.4 순환신경망(LSTM)의 RMSE ····· 15
Table 3.5 양방향 순환신경망의 RMSE ····· 15
Table 3.6 상승/하락 예측 ····· 22
Table 3.7 추세전환 시뮬레이션을 위한 모델 ····· 23
Table 3.8 LSTM 추세전환 시뮬레이션 ····· 23
Table 3.9 양방향 순환신경망 추세전환 시뮬레이션 ····· 23

그림목차

Fig. 2.1 인공신경망 ····· 5
Fig. 2.2 시그모이드 함수 ····· 6
Fig. 2.3 오차역전파 ····· 6
Fig. 2.4 순환신경망 구조 ····· 8
Fig. 2.5 순환신경망 전체 구조 ····· 8
Fig. 2.6 LSTM 구조 ····· 9
Fig. 2.7 양방향 순환신경망 구조 ····· 11
Fig. 3.1 신경망 계층 비교 ····· 13
Fig. 3.2 Case 1 ····· 16
Fig. 3.3 Case 2 ····· 16
Fig. 3.4 Case 3 ····· 17
Fig. 3.5 Case 4 ····· 17
Fig. 3.6 Case 5 ····· 17
Fig. 3.7 Case 6 ····· 18

Fig. 3.8 Case 7 ·····18
Fig. 3.9 Case 8 ·····18
Fig. 3.10 Case 9 ·····19
Fig. 3.11 Case 10 ·····19
Fig. 3.12 Case 11 ·····19
Fig. 3.13 Case 12 ·····20
Fig. 3.14 Case 13 ·····20
Fig. 3.15 Case 14 ·····20
Fig. 3.16 Case 15 ·····21
Fig. 3.17 Case 16 ·····21

Ⅰ. 서 론

1. 연구의 배경 및 목적

시계열 데이터(Time-series Data)는 주식, 매출액, 실업률 등과 같이 시간적 순서를 가지는 데이터를 의미한다. 현재 직면한 문제를 해결하려는 모든 조직에게 시계열 데이터의 관찰, 분석 더 나아가 데이터 예측은 필수다. 데이터 예측을 통해 조직 내외 의사결정 문제를 해결할 수 있을 뿐만 아니라 의사결정의 결과로서의 계획, 투자, 성과를 향상시킬 수 있으며 투자 기회비용, 재고비용, 물류비용 등과 같은 비용의 절감을 할 수 있기 때문이다.

시계열 데이터를 다루는 전통적인 수학 모델의 한계를 극복하기 위한 대안으로 인공신경망을 사용하는 기계학습 기법이 제안되었다. 인공신경망의 일종인 순환신경망(Recurrent Neural Networks)은 비선형적 분류가 필요한 시계열 데이터의 분석에 적합하다. 순환신경망(RNN)은 일반 인공신경망에 시계열 개념을 추가한 것으로, 은닉계층에 이전 정보를 기억시킬 수 있다. 그러나 데이터양과 길이가 늘어나면서 과거의 학습한 결과가 사라지는 장기 의존성 문제 때문에 성능이 하락하는 단점을 갖고 있다. 이 문제를 해결하기 위해 LSTM(Long Short-Term Memory)이 제안되었는데, LSTM은 전체 체인을 관통하는 셀 상태(Cell State)를 통해 과거 학습결과를 전달하는 구조를 갖는다. 이러한 셀 상태를 유지하면서 입력게이트, 망각게이트와 출력게이트를 이용하여 출력 값을 조정해나가는 방식으로 기존의 장기 의존성 문제를 극복할 수 있다. 그러나 LSTM은 입력 데이터를 시간 순서대로 처리하기 때문에 결과가 이전 시점의 데이터에 영향을 많이 받는다. 이러한 점 때문에 여러 가지 요인이 영향을 주는 데이터와 같은 경우에는 계산상의 손실이 크게 발생한다. 이를 극복하기 위해서 순방향과 역방향으로 분리된 순환신경망을 통해 데이터를 학습시키는 양방향 순환신경망(Bidirectional Recurrent Networks)이 제안되었다[1][2]. 이는 미래의 데이터를 고려하기 때문에 순방향 순환신경망보다 성능이 우수하다는 장점을 가지고 있다.

본 논문에서는 시계열 데이터 중 주식 예측모델을 다룬다. 기존에 순환신경망, LSTM과 양방향 순환신경망을 이용한 주식 예측모델에 대한 연구가 있었지만, 가격과 거래량 데이터만을 사용하였다. 본 논문에서는 가격, 거래량 데이터에 가격 보조지표, 거래량 보조지표, 추세 보조지표와 채널 보조지표를 추가한 모델들 간의 비교를 통해 어떤 데이터를 사용하는 것이 주식을 예측하는데 적절한지를 평가한다.

2. 논문의 구성

본 논문의 구성은 다음과 같다. 제1장 서론에서는 연구 배경을 제2장 관련 연구에서는 기술적 분석 방법, 인공신경망, 순환신경망, Long Short-Term Memory 그리고 양방향 순환신경망에 대해 설명하며 제3장에서는 모델의 실험과 성능 평가, 성능 검증에 대해 설명한다. 제4장에서는 결론을 기술한다.

Ⅱ. 관련 연구

1. 기술적 분석

기술적 분석은 주가가 많은 복합적인 요소들을 반영하는 결과라는 가정에서 출발하여 주가 그 자체를 분석함으로써 미래의 주가 변동을 예측하려는 분석 방법이다. 기술적 분석은 주가, 거래량과 보조지표를 가지고 상태를 평가한 후 투자를 결정하려는 분석법이다[3][4]. 보통 보조지표들로서는 다음 Table 2.1 ~ Table 2.4의 가격, 거래량, 추세, 채널 보조지표들이 활용된다.

Table 2.1 가격 보조지표

Category	Index	설명
Price	가격 이동평균	일정기간 주가의 평균 가격을 의미한다. 특정기간 동안의 산술평균을 말하는 것으로 5일 단순이동 평균은 지난 5일 동안 가격의 평균을 의미한다.
	Box Chart	일정개수로 설정된 묶음 수 캔들의 첫 번째 봉의 시가, 묶음 캔들 중의 고가, 묶음 캔들 중의 저가, 마지막 종가를 데이터 기준으로 한 개의 캔들 차트 형태로 보여주는 지표이다. 5일씩 캔들을 묶어서 5일 Box를 만든 후 Box의 시가, 고가, 저가, 종가를 나타낸다.
	Weighted Close	Weighted Close는 매일 매일 주가를 단순 평가한 종가를 강조하는 지표이다.

Table 2.2 거래량 보조지표

Category	Index	설명
Volume	거래량 이동평균	매입수요가 상대적으로 증가할 때 주가가 상승하며 이에 따라 거래량이 늘어나는 경향을 보이고, 반대로 매입수요가 감소하여 주가가 하락할 때에는 거래량이 줄어들게 된다. 거래량 이동평균에는 단기지표로 5일, 중기지표로 20일, 장기지표로 60일을 많이 활용한다.
	Daily Volume Index	당일의 매수와 매도 세력을 고려하여 만든 거래량 지표의 일종이다.
	VR (Volume Ratio)	일정 기간 동안(20일) 시장가격 상승일의 거래량과 시장가격 하락일의 거래량을 비교하여 백분율로 나타낸 지표로서 현재시장이 과열인지 침체인지를 판단하게 해주는 시장특성 지표이다.

Table 2.3 추세 보조지표

Category	Index	설명
Trend	RSI (Relative Strength Index)	RSI는 시장가격의 변동 폭 중에서 상승폭이 어느 정도인지를 분석하여, 현재의 시장가격이 상승세라면 얼마나 강력한 상승추세인지, 그리고 하락세라면 얼마나 강력한 하락추세인지를 백분율(Percentage)로 나타낸 것이다. 추세의 강도를 표시하여 주므로 향후 추세전환시점 예측이 가능한 지표다.
	RSI+MACD	MACD는 단기 지수 이동 평균 값에서 장기 지수 이동 평균값을 뺀 차이로 두 이동 평균 사이의 관계를 보여 주는 지표다. RSI는 현재 추세의 강도를 백분율로 나타내 언제 주가 추세가 전환될 것인가를 예측하는데 유용한 지표다. 이 두 지표를 혼합한 형태가 RSI+MACD다.
	RSI+Stochastic	RSI는 언제 주가 추세가 전환될 것인가를 예측하는데 유용한 지표다. Stochastic은 일정 기간 동안의 주가 변동 폭 중에서 금일 종가의 위치를 백분율로 나타내는 지표로서 주가의 움직임을 반영하는 지표다. 이 두 지표를 혼합한 형태가 RSI+Stochastic이다.

Table 2.4 채널 보조지표

Category	Index	설명
Channel	Bollinger Bands	Bollinger Bands의 상, 하한선은 표준 편차에 의해 산출된 이동평균 값이며 주가나 지수의 움직임이 큰 시기에는 Bands의 폭이 넓어지고 움직임이 작은 시기에는 Bands의 폭이 좁아지는 특성을 가지고 있다. 가격움직임의 크기에 따라 밴드의 넓이가 결정된다.
	Parabolic SAR (Stop and Reversal)	Parabolic SAR는 추세전환시점을 알려주는 지표로써 Welles Wilder에 의해 개발된 지표다.

2. 인공신경망

인공신경망은 인간의 뇌 구조를 컴퓨터로 구현하여 보다 분류(classification)에 적합한 모델이다[1]. 특히, 비선형 분석모델(Nonlinear analytical model)로서, 의사결정에 있어서 인간의 두뇌를 모방하려는 시도에서 개발되었다. 인공신경망은 입력계층과 출력계층 사이에 뉴런을 추가하여 스스로 가중치 값을 설정하고 조정해나가며 자동으로 학습하는 알고리즘이다. Fig. 2.1과 같이 인공신경망은 입력계층, 은닉계층, 출력계층으로 구성되어 있으며, 은닉계층은 눈으로 확인할 수 없는 계층으로 가중치를 합산한다.

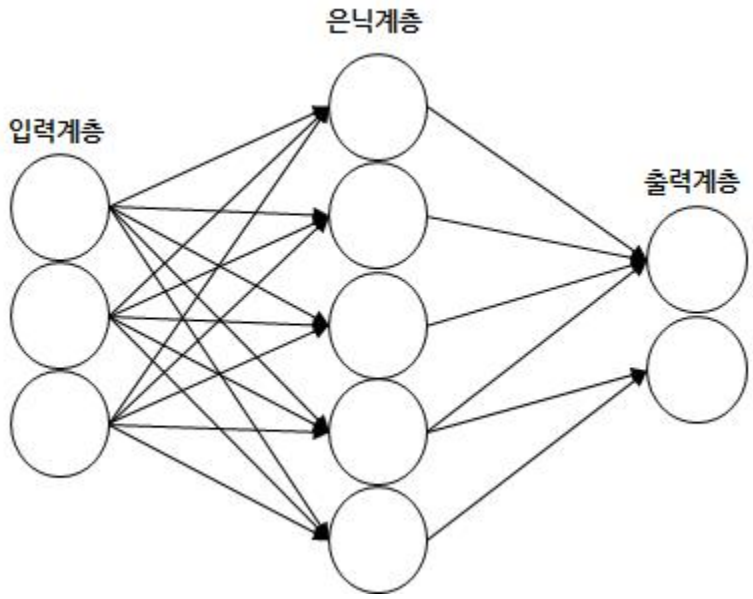

Fig. 2.1 인공신경망

1) 활성화 함수

인공신경망은 은닉 계층에서 가중치를 합산할 때 활성화 함수(Activation Function)를 통해 변환하여 출력계층으로 보내는 구조를 가지고 있다[1]. 활성화함수는 출력계층의 값이 커지면 최종 출력이 1에 수렴하는 함수로 이를 통해 비선형 자료(Nonlinear Data)에서 우수한 성능을 가질 수 있다.

활성화 함수로는 주로 S자 함수가 사용되며, 로지스틱 함수와 쌍곡 탄젠트 함수가 사용된다. 활성화 함수의 종류에는 Fig. 2.2와 같은 로지스틱 함수인 시그모이드 함수(Sigmoid), 쌍곡 탄젠트 함수 그리고 정류 선형 유닛 함수(Rectified Linear Unit, ReLU) 등이 있다.

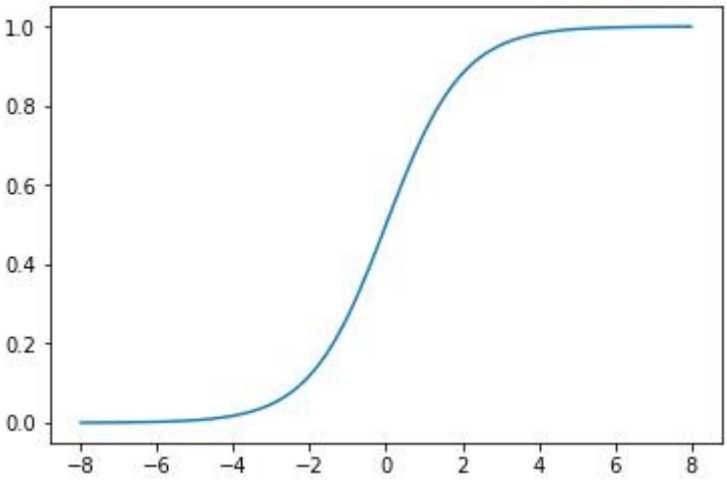

Fig. 2.2 시그모이드 함수

2) 오차 역전파

Fig. 2.3과 같이 입력에서 출력으로 가중치를 갱신하면서 활성화 함수를 거쳐 순전파 출력 값(y)을 산출하는 과정이 순전파(Forward) 과정이다[1]. 이때 오차가 발생하게 되는데 최종 출력 값(o)과 순전파 출력 값(y) 사이에 ($y-o$) 만큼의 오차(e)가 발생한다. 그래서 오차를 줄이기 위해 오차 역전파는 Fig. 2.3과 같이 역방향으로 오차(e)를 다시 보내며 가중치를 갱신한다. 순전파 과정과 역전파 과정을 오차의 값이 최소가 될 때까지 반복함으로써 최적의 가중치를 산출하게 된다.

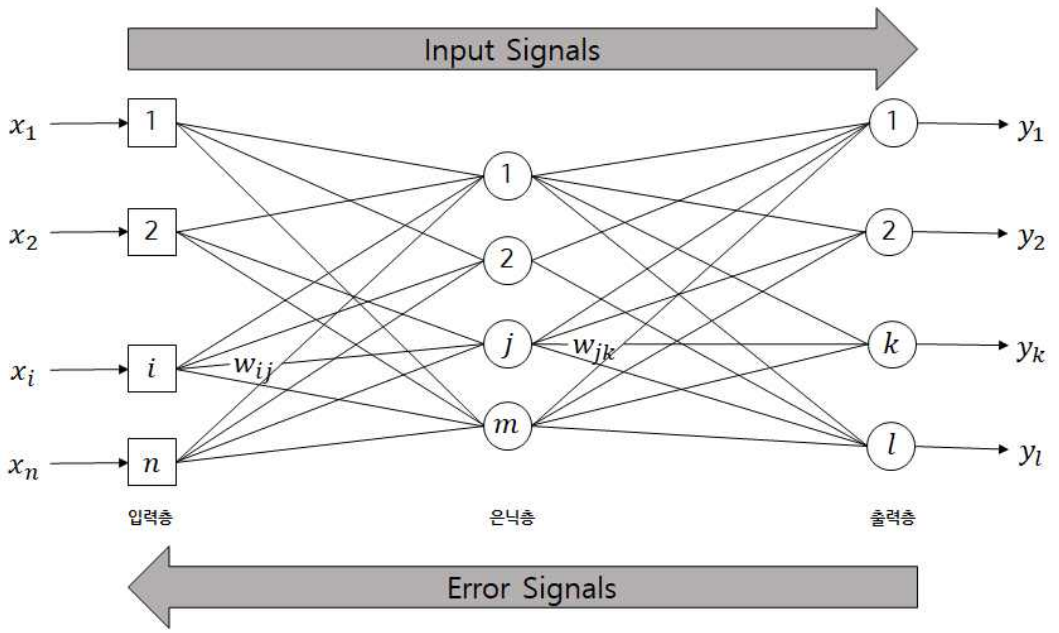

Fig. 2.3 오차 역전파

3. 순환신경망(Recurrent Neural Networks)

인공신경망 모형 중 순환신경망(Recurrent Neural Network)은 일반 신경망에 시계열 개념을 추가한 것으로, 은닉계층에 이전 정보를 기억시킬 수 있는 장점이 있다[1]. 본 논문은 순환신경망을 통해 과거의 지수 가격이 미래의 지수 가격에 어떻게 영향을 미치는지를 예측한다. 순환신경망은 은닉계층에 과거의 지수 데이터를 기억하여 학습을 진행하기 때문에 시계열 데이터인 Kospi 200 지수 예측에 적합한 모형이다.

1) 순환신경망의 기본 구조

순환신경망은 동일한 작업을 입력되는 데이터에 순서적으로 적용하고, 이전 출력 결과의 영향을 받도록 현재까지 계산된 결과를 메모리에 기억한다[1].

Fig. 2.4와 같이 순환신경망 구조에서는 입력 값을 x_t로 받고 출력 값 h_t를 출력한다. 순환신경망은 Fig. 2.4와 같이 동일한 네트워크를 여러 개 반복하는 것과 같이 루프로 연결되어 있다. 이러한 순환신경망 구조를 펼치면 Fig. 2.5와 같다. 순환신경망의 식 (2.1)에서 x_t는 입력데이터, h_{t-1}은 이전 단계의 은닉계층 출력, W는 가중치, b는 바이어스, H는 활성화 함수를 나타낸다[6]. t 단계의 은닉계층 출력 h_t는 입력데이터에 가중치를 곱하고 이전 단계의 은닉계층 출력에 가중치를 곱한 값과 바이어스를 더한 후 활성화 함수를 적용하여 산출된다. 이를 통해 순환신경망은 과거의 데이터가 다음 데이터에 어떤 영향을 주는지 학습할 수 있고, 시계열 데이터인 지수 데이터 예측에 사용할 수 있다. 그 외에도 순환신경망은 음성인식, 영상인식, 문장번역, 자연어처리 등에 활용된다.

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2.1}$$

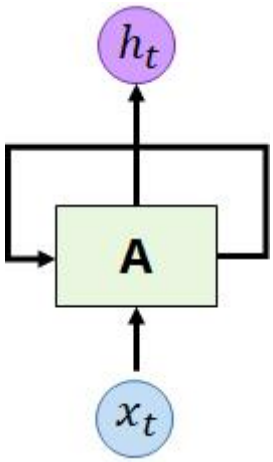

Fig. 2.4 순환신경망 구조

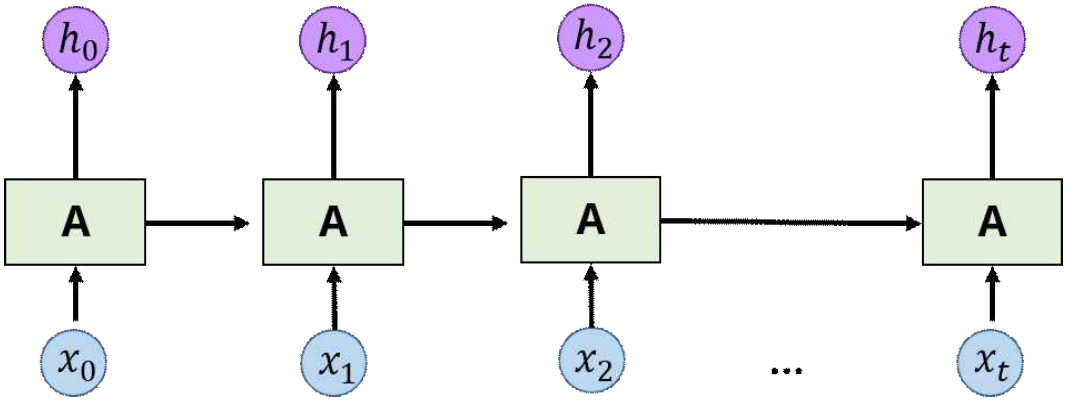

Fig. 2.5 순환신경망 전체 구조

(2) 장기의존성 문제

순환신경망의 학습은 입력을 비선형 관계로 변환하는 활성화함수인 Sigmoid, tanh 함수를 이용하여 출력을 하며, 이 값을 다음 단계의 입력으로 전달한다. 출력 값은 -1 ~ 1사이의 작은 값을 갖기 때문에 출력 값이 곱해지는 단계가 반복되면 출력값은 작은 값을 갖게 된다. 이러한 절차가 계속되면서 출력 값이 소실되어 학습이 더 이상 이루어지지 않는 기울기 소실(Vanishing gradient)이 발생한다. 다시 말하면, 과거의 학습의 결과가 사라지는 장기의존성(Long-Term dependencies) 문제가 발생한다[1].

이러한 기울기 소실 문제를 해결하기 위해 LSTM(Long-Short Term Memory) 구조를 사용하며 LSTM을 통해 긴 시퀀스(Sequence)를 효과적으로 처리할 수 있다.

4. Long Short-Term Memory

LSTM(Long-Short Term Memory)은 전체 체인을 관통하는 셀 상태(Cell State)를 통해 과거 학습결과를 전달하는 구조로서 장기 의존성 문제를 해결할 수 있는 순환신경망 구조이다[1].

(1) LSTM 구조

LSTM 셀은 셀 상태를 유지하면서 입력게이트, 망각게이트와 출력게이트를 이용하여 출력 값을 조정한다. Fig. 2.6과 같이 LSTM의 셀 상태는 게이트를 이용하여 셀 상태를 제어하고 정보의 반영여부를 정한다. 입력게이트(Input Gate)는 식 (2.2)에 의해 입력 값을 얼마나 받아들일 것인가를 정하고, 망각게이트(Forget Gate)는 식 (2.3)에 의해 이전의 셀 상태를 얼마나 잊어버릴지를 정하며, 출력게이트(Output Gate)는 식 (2.5)에 의해 무엇을 얼마나 출력할 것인가를 정한다.

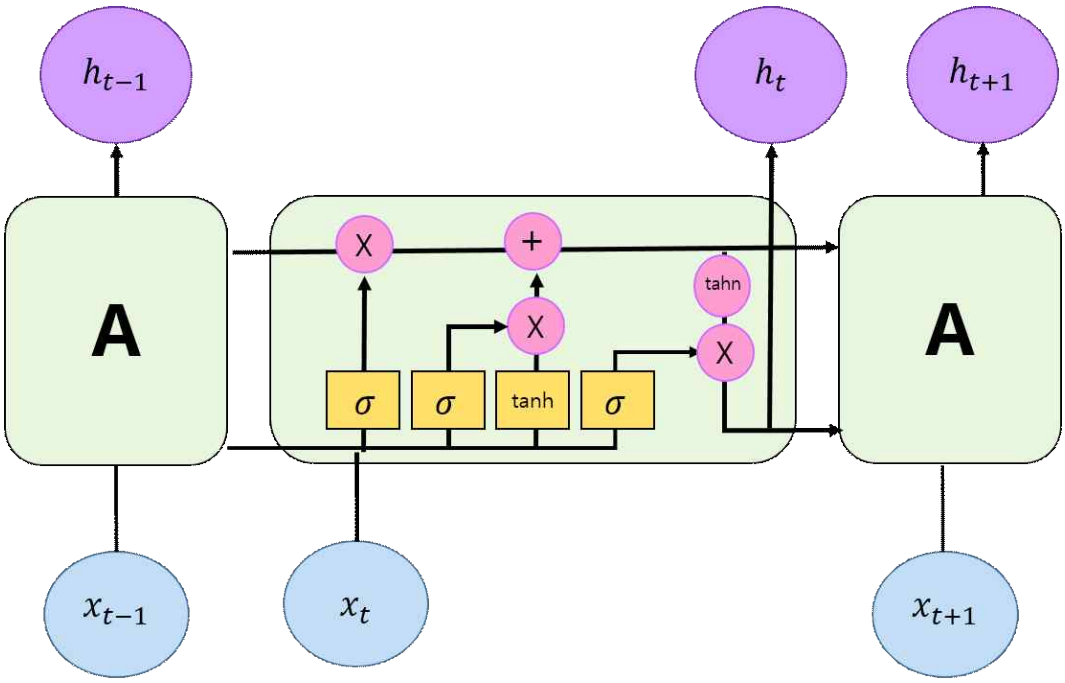

Fig. 2.6 LSTM 구조

$$i_t = \sigma(W_\xi x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \quad (2.2)$$

$$f_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f) \quad (2.3)$$

$$c_t = f_t c_{t-1} + i_t Tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \quad (2.4)$$

$$O_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \quad (2.5)$$

$$h_t = o_t Tanh(c_t) \quad (2.6)$$

위에서 c_t는 셀 상태를 나타내며, h_t는 셀의 출력이다. i_t, f_t, o_t는 각각 입력게이트, 망각게이트, 출력게이트를 나타내며, σ와 $Tanh$은 활성화 함수로 각각 시그모이드(Sigmoid)와 하이퍼탄젠트(Hyperbolic Tangent)를 나타낸다.

식 (2.4)는 기존의 셀 상태에 망각게이트의 출력 값을 곱하여 셀 상태의 일정기간의 값을 잊어버리고 입력 값과 이전 단계의 출력 값을 처리한 결과에 입력게이트의 출력 값을 곱하여 입력을 받으면서 새로운 셀 상태를 만드는 것을 보여준다. 또한 식 (2.6)은 셀의 출력이 이 셀 상태에 출력게이트의 출력 값을 곱해서 얻어지는 것을 보여준다. 위와 같이 셀 상태 값을 얼마나 잊어버리고 새로운 입력 값을 얼마나 받아들이지를 정할 수 있기 때문에 장기의존성 문제를 해결할 수 있다.

5. 양방향 순환신경망(Bidirectional Recurrent Neural Networks)

양방향 순환신경망(Bidirectional Recurrent Neural Networks)은 특정시점의 출력 값이 이전 시점과 이후 시점의 데이터까지 고려하는 RNN의 확장 모델이다[1][2]. RNN은 입력 순서를 시간 순서대로 처리하기 때문에 결과는 이전 시점의 데이터에 영향을 많이 받는 단점을 가지고 있다. 이러한 단점을 극복하기 위해서 순방향과 역방향으로 분리된 순환신경망을 통해 학습시킨다. 미래의 데이터를 고려하기 때문에 순방향 순환신경망보다 성능이 우수하다.

(1) 양방향 순환신경망 구조

Fig. 2.7과 같이 양방향 순환신경망은 순방향 은닉계층과 역방향 은닉계층을 가지고 있으며, 서로 연결되지 않는다. 입력 값은 두 개의 은닉계층으로 전달되고 식 (2.7), (2.8)과 같이 순방향 은닉계층 출력 a_t^f와 역방향 은닉계층 출력 a_t^b가 산출된다. 출력계층에서는 식 (2.9)에 의해 두 개의 은닉계층의 출력을 하나로 만들어 입력을 받은 후, 가중치를 계산하여 최종 출력을 한다.

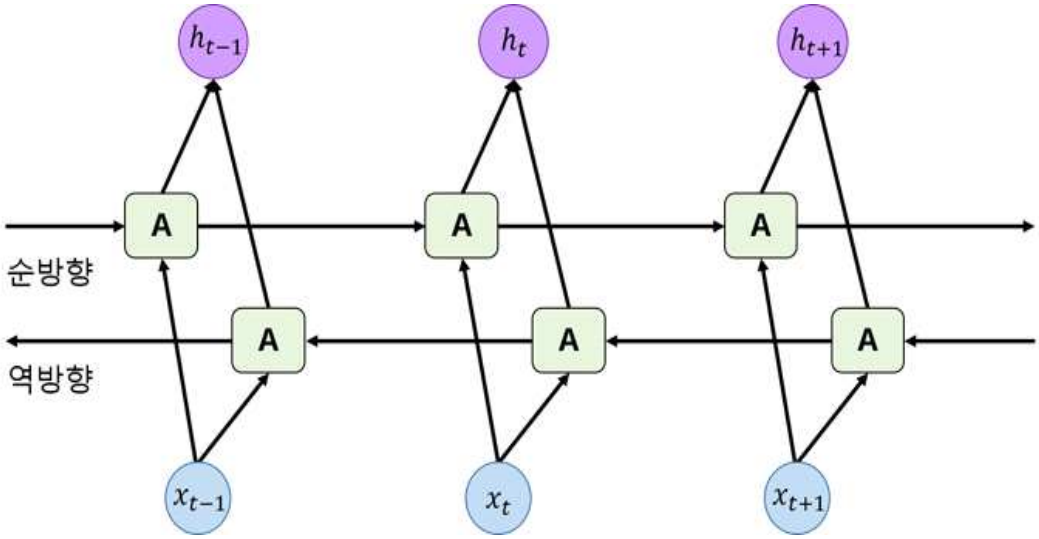

Fig. 2.7 양방향 순환신경망 구조

$$a_t^f = \Phi(W_{a^f} x_t + W_{a^f} a_{t-1}^f + b_{a^f}) \tag{2.7}$$

$$a_t^b = \Phi(W_{a^b} x_t + W_{a^b} a_{t-1}^b + b_{a^b}) \tag{2.8}$$

$$h_t = W_{a^f} a_t^f + W_{a^b} a_t^b + b_y \tag{2.9}$$

Ⅲ. 실험 및 성능평가

1. 실험 환경

본 논문에서 사용된 주식 가격 데이터는 2000년 01월 04일부터 2018년 08월 31일까지의 Kospi 200 지수인 4612개의 데이터를 사용하였다. 분석 데이터는 일자별 주식의 시가, 고가, 저가, 종가, 거래량, 가격 이동평균(5일, 20일 60일), 거래량 이동평균(5일, 20일, 60일), Box Chart, Weighted Close, Bollinger Bands, Parabolic SAR, Daily Volume Index, VR, RSI, RSI+MACD, RSI+Stochastic들을 사용하였다. 전체 주식 데이터에서 3612개의 데이터를 학습용 데이터로, 나머지 1000개를 테스트용 데이터로 선택하여 실험을 진행했다. 지수 가격 예측의 오차율을 확인하기 위해 식 (3.1)의 실제 지수 가격과 예측 지수 가격간의 평균 제곱근 오차(Root Mean Square Error)를 사용했다.

$$\text{RMSE} = \sqrt{\frac{1}{n}\left(\sum_{t=1}^{n}(\text{실제 지수가격의 종가}_t - \text{예측지수가격}_t)^2\right)} \qquad (3.1)$$

본 논문에서는 지수 가격을 예측하기 위해 인공신경망 모형인 LSTM과 양방향 순환신경망을 사용했다. 네트워크 구조는 시가, 고가, 저가, 종가, 거래량, 가격 보조지표, 채널 보조지표, 거래량 보조지표, 추세 보조지표를 입력 계층으로 구성하여 LSTM 셀로 구성했다. Fig. 3.1에서 보는 바와 같이 성능이 가장 좋은 3개의 은닉계층을 선택하였으며, Table 3.1과 같이 Layer 비교를 통해 비교적 시간이 적게 걸리는 Epoch 10개를 기준으로 선택하였다. 즉, 본 논문의 순환신경망 계층은 1개의 입력 계층, 3개의 은닉계층, 1개의 출력 계층으로 이루어진다. 순환신경망의 활성화 함수로는 시그모이드(Sigmoid) 함수를 사용했으며, 순환신경망의 최적화 방식은 아담 최적화기(Adam Optimizer)를 이용하여 과거 및 현재의 지수 가격, 거래량과 보조지표를 이용하여 미래의 지수 가격을 학습한다. 본 논문의 실험환경은 Table 3.2와 같다.

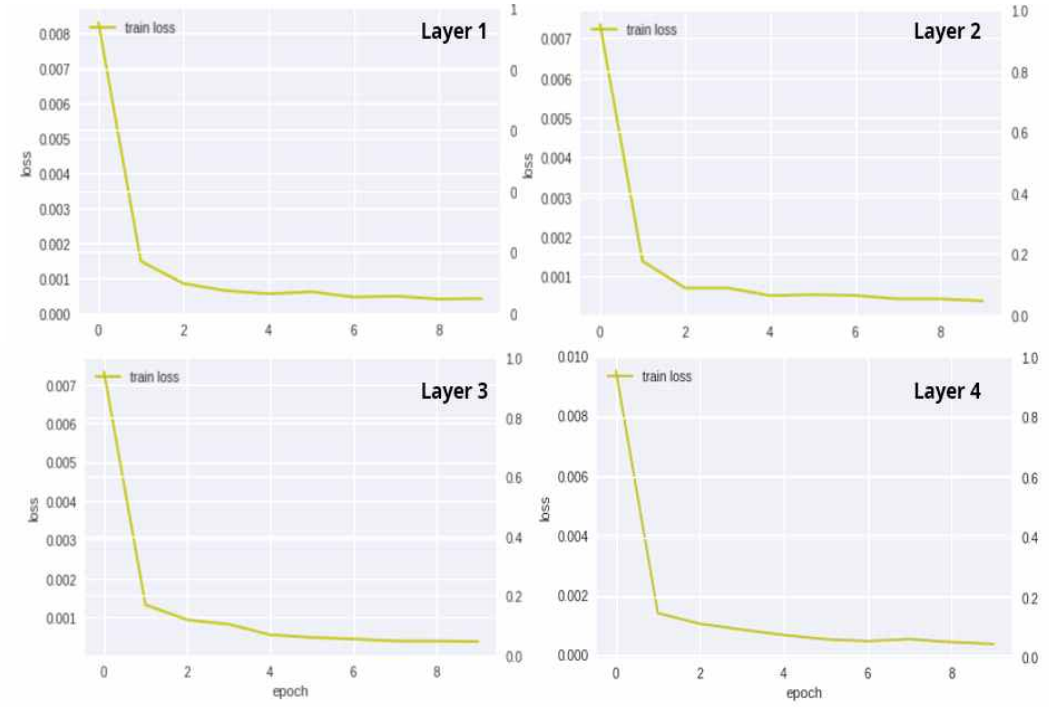

Fig. 3.1 양방향 순환신경망 계층 비교

Table 3.1 양방향 순환신경망 계층 비교

양방향 순환신경망 기준	Epoch 10번
Layer 1	17.17초
Layer 2	37.80초
Layer 3	38.39초
Layer 4	40.61초

Table 3.2 실험환경

구성 요소	규격
OS	Windows 10(64-Bit)
CPU	Intel(R) Core(TM) i5-3550
RAM	12GM
Python	3.6.4
Keras	2.2.2

2. 성능 평가

본 실험은 오픈소스 인공신경망 라이브러리인 케라스(Keras)를 이용하여 실험을 하였다. 학습률은 0.01로 설정하고, 500번의 반복 학습을 수행하였다. 본 논문에서 선정된 순환신경망 모델에 Table 3.3과 같이 보조지표의 서로 다른 조합을 갖는 Case들을 고려하고 각각의 RMSE를 측정하였다. 그리고 전날 종가와 예측 값의 비교를 통해 Kospi 지수의 상승/하락률을 계산하였다.

Table 3.3 Case 분류

보조지표	Case
가격, 거래량	Case1
가격, 거래량, 가격보조지표	Case2
가격, 거래량, 거래량 보조지표	Case3
가격, 거래량, 추세 보조지표	Case4
가격, 거래량, 채널 보조지표	Case5
가격, 거래량, 가격 보조지표, 거래량 보조지표	Case6
가격, 거래량, 가격 보조지표, 추세 보조지표	Case7
가격, 거래량, 가격 보조지표, 채널 보조지표	Case8
가격, 거래량, 거래량 보조지표, 추세 보조지표	Case9
가격, 거래량, 거래량 보조지표, 채널 보조지표	Case10
가격, 거래량, 추세 보조지표, 채널 보조지표	Case11
가격,거래량,가격보조지표,거래량보조지표,추세보조지표	Case12
가격, 거래량, 가격 보조지표, 거래량 보조지표, 채널 보조지표	Case13
가격, 거래량, 가격 보조지표, 추세 보조지표, 채널 보조지표	Case14
가격, 거래량, 거래량 보조지표, 추세 보조지표, 채널 보조지표	Case15
가격, 거래량,가격보조지표,거래량보조지표,추세보조지표,채널보조지표	Case16

Table 3.4 순환신경망(LSTM)의 RMSE

Case	Max	Min
Case1	0.037493	0.024837
Case2	0.02919	0.017617
Case3	0.041865	0.031956
Case4	0.038779	0.029763
Case5	0.052261	0.034981
Case6	0.055496	0.037135
Case7	0.034803	0.018722
Case8	0.030204	0.019803
Case9	0.062191	0.046972
Case10	0.069139	0.053207
Case11	0.047847	0.039991
Case12	0.049364	0.032827
Case13	0.049026	0.029636
Case14	0.030708	0.021494
Case15	0.079839	0.050138
Case16	0.054905	0.035257

Table 3.5 양방향 순환신경망의 RMSE

Case	Max	Min
Case1	0.042986	0.024145
Case2	0.023516	0.016356
Case3	0.05456	0.043693
Case4	0.028661	0.024125
Case5	0.056247	0.03017
Case6	0.05126	0.030595
Case7	0.027248	0.01866
Case8	0.024571	0.016518
Case9	0.051476	0.039979
Case10	0.051078	0.042858
Case11	0.042653	0.030797
Case12	0.045345	0.02843
Case13	0.039391	0.026389
Case14	0.033715	0.020809
Case15	0.057319	0.030459
Case16	0.04486	0.026551

Table 3.4와 Table 3.5는 순환신경망과 양방향 순환신경망을 사용했을 때 각 Case들의 RMSE를 나타낸다. Table 3.4에서 가장 좋은 성능을 낸 LSTM 모델은 입력 데이터로서 가격, 거래량, 가격보조지표를 사용한 Case2로서 최대값과 최소값이 각각 0.02919, 0.017617이었다. Table 3.5에서 가장 좋은 성능을 낸 양방향 순환신경망은 역시 가격, 거래량, 가격보조지표를 사용한 Case2로

서 최대값과 최소값이 각각 0.023516, 0.016356이었다. 또한 가격, 거래량, 가격보조지표를 사용한 Case2에서 양방향 순환신경망이 LSTM보다 좋은 성능을 보였다. Table 3.4~Table 3.5, Fig.3.2~Fig.3.17로부터 RMSE 기준으로 16개의 Case 중에서 15개에서 양방향 순환신경망이 LSTM보다 우수함을 볼 수 있었다. 그리고 가장 성능이 좋은 모델은 Case2, Case8, Case7, Case14 순이었다. 이들이 갖고 있는 보조지표들은 가격 보조지표, 추세 보조지표, 채널 보조지표이다. 즉, 모델 선정 시 가격 보조지표, 추세 보조지표, 채널 보조지표를 쓴다면 모델의 정확성이 높아질 것이다.

1) 가격, 거래량

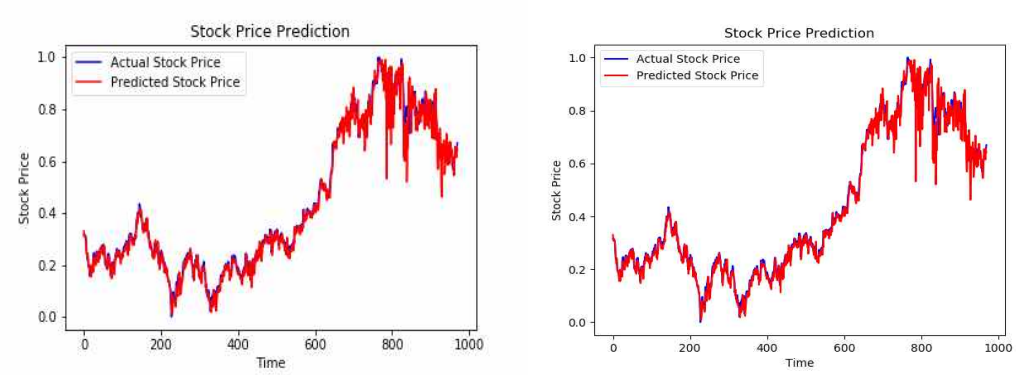

Fig. 3.2 Case 1

2) 가격, 거래량, 가격 보조지표

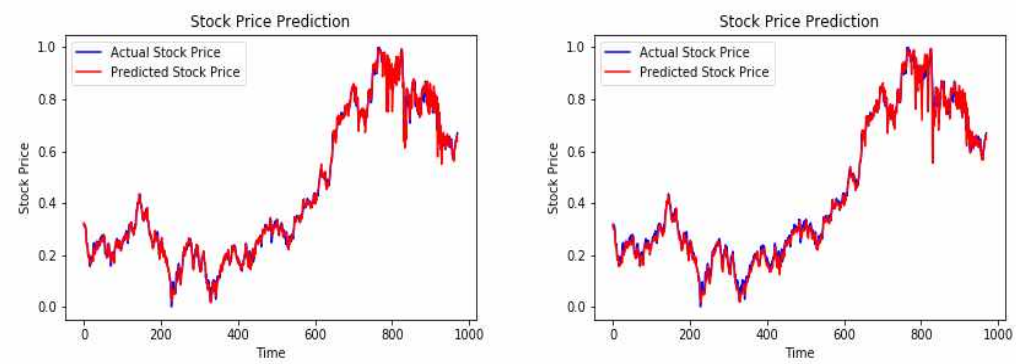

Fig. 3.3 Case 2

3) 가격, 거래량, 거래량 보조지표

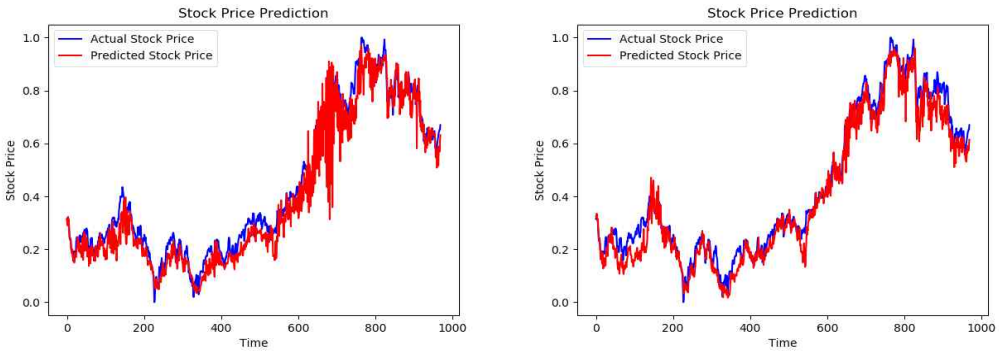

Fig. 3.4 Case 3

4) 가격, 거래량, 추세 보조지표

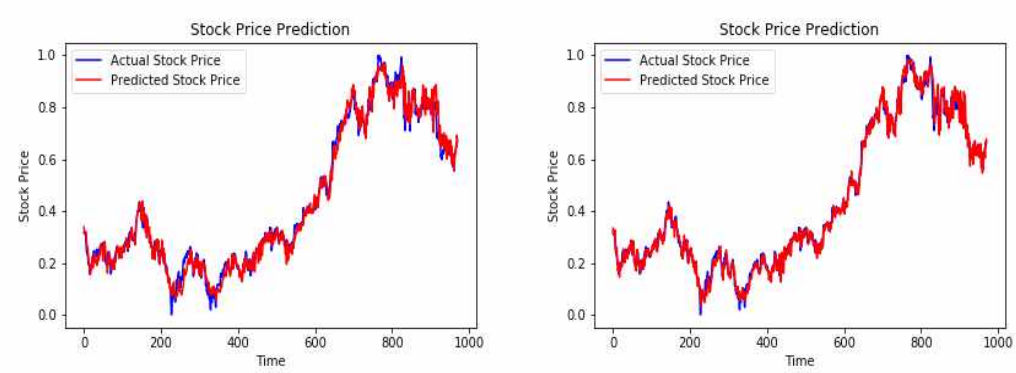

Fig. 3.5 Case 4

5) 가격, 거래량, 채널 보조지표

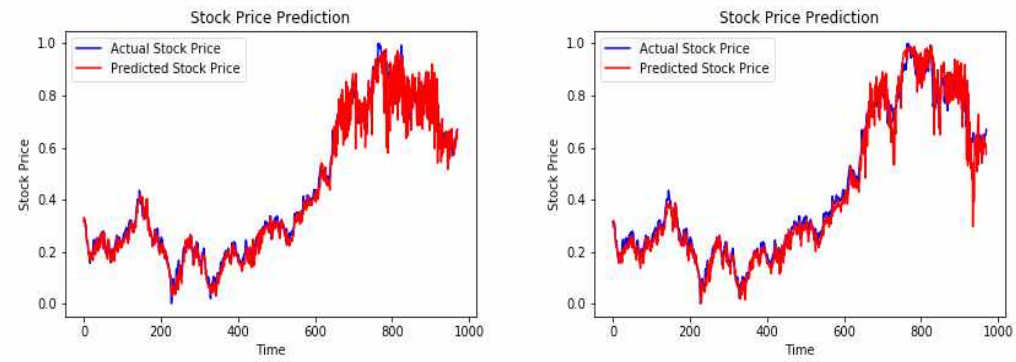

Fig. 3.6 Case 5

6) 가격, 거래량, 가격 보조지표, 거래량 보조지표

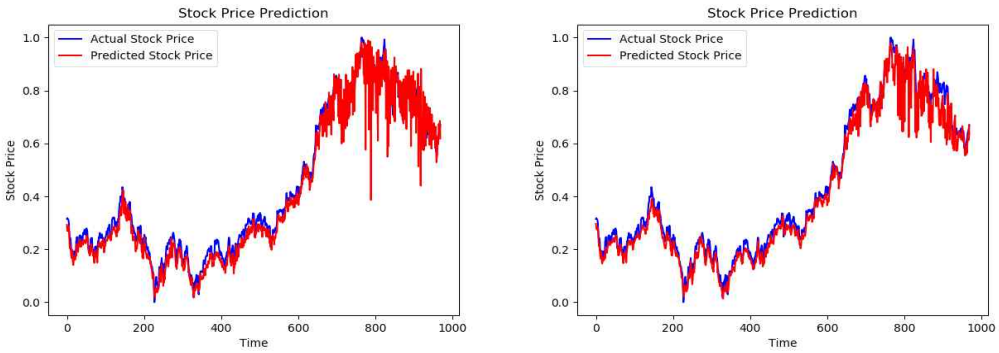

Fig. 3.7 Case 6

7) 가격, 거래량, 가격 보조지표, 추세 보조지표

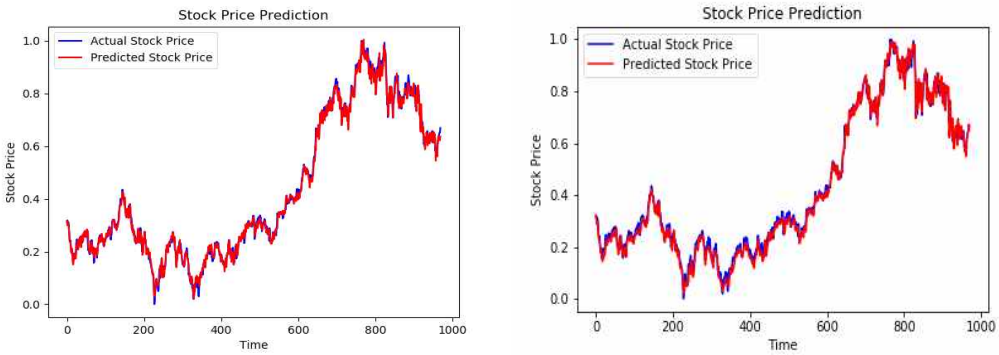

Fig. 3.8 Case 7

8) 가격, 거래량, 가격 보조지표, 채널 보조지표

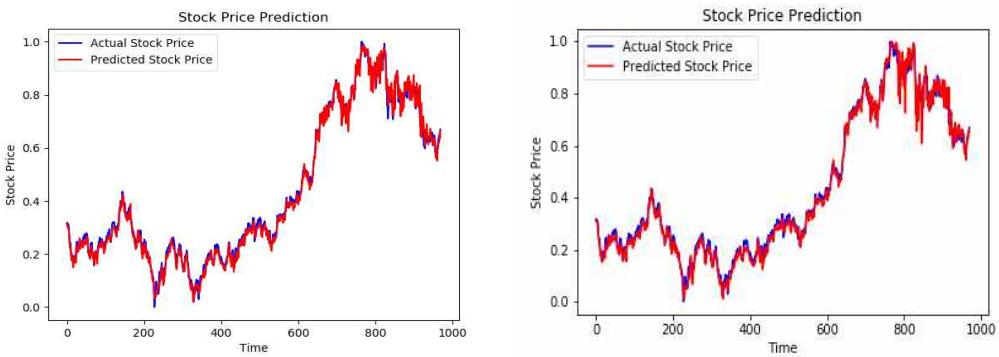

Fig. 3.9 Case 8

9) 가격, 거래량, 거래량 보조지표, 추세 보조지표

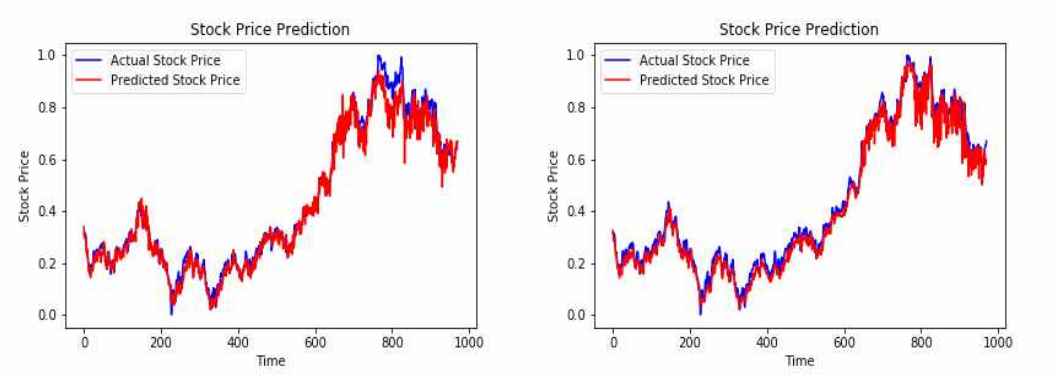

Fig. 3.10 Case 9

10) 가격, 거래량, 거래량 보조지표, 채널 보조지표

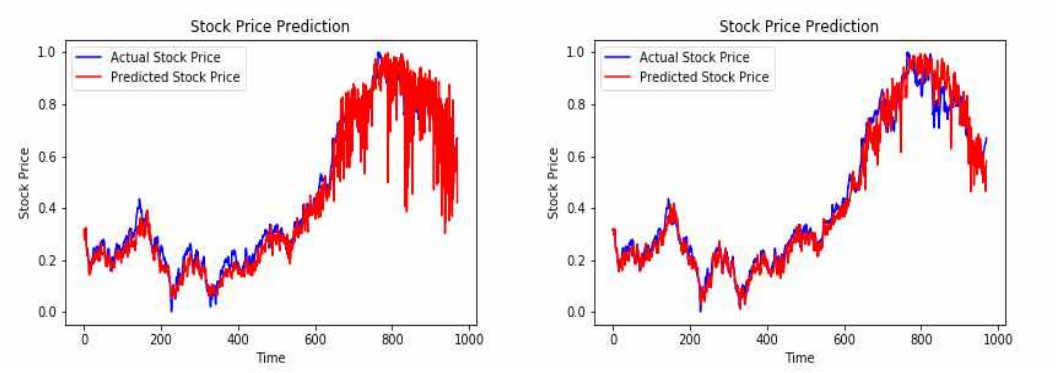

Fig. 3.11 Case 10

11) 가격, 거래량, 추세 보조지표, 채널 보조지표

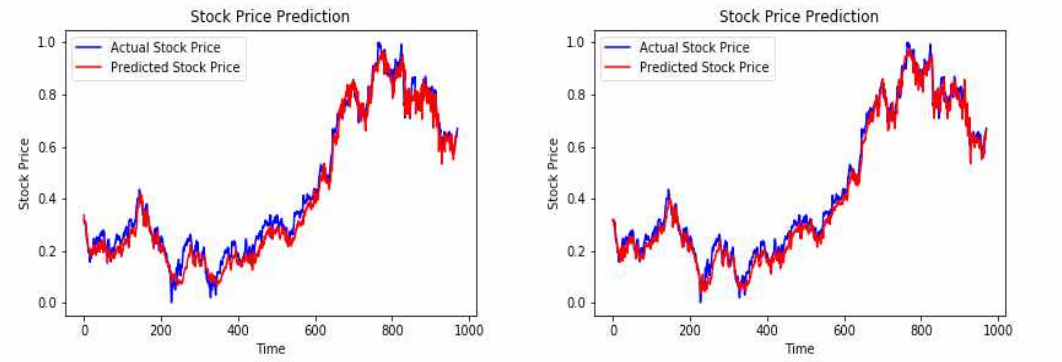

Fig. 3.12 Case 11

12) 가격, 거래량, 가격 보조지표, 거래량 보조지표, 추세 보조지표

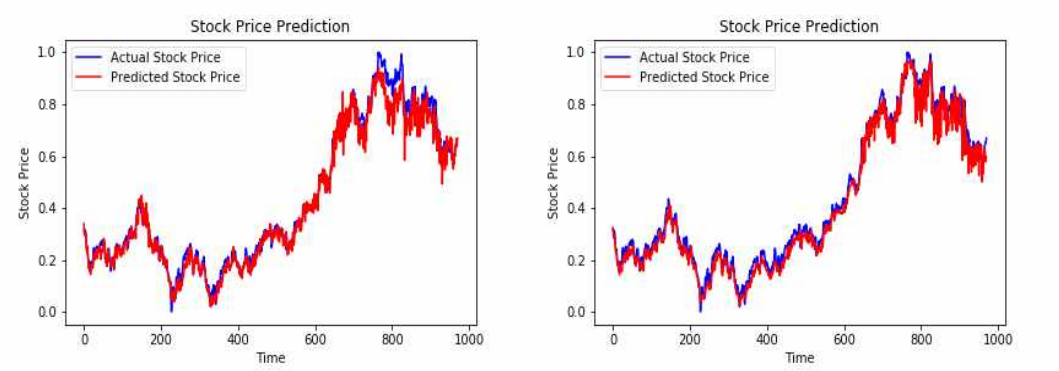

Fig. 3.13 Case 12

13) 가격, 거래량, 가격 보조지표, 거래량 보조지표, 채널 보조지표

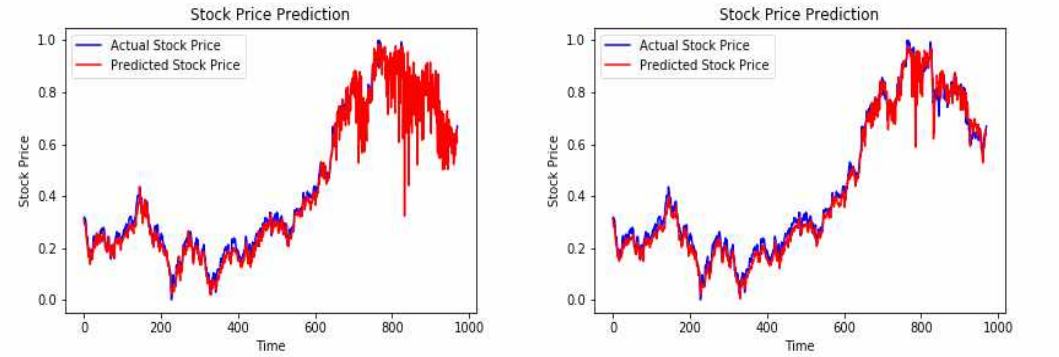

Fig. 3.14 Case 13

14) 가격, 거래량, 가격 보조지표, 추세 보조지표, 채널 보조지표

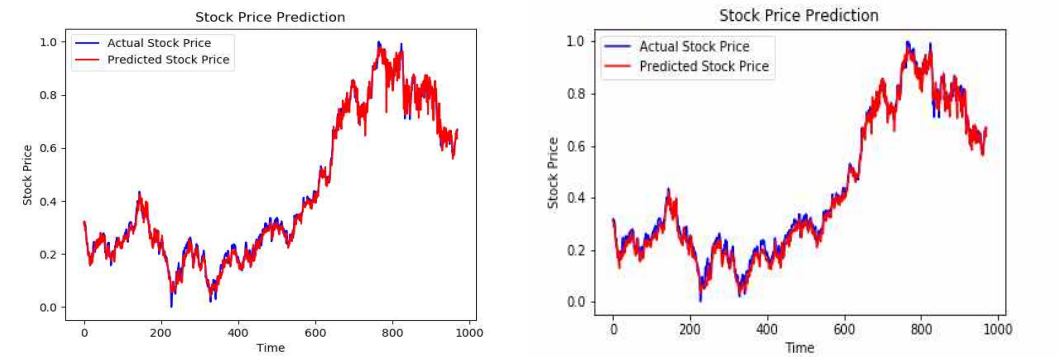

Fig. 3.15 Case 14

15) 가격, 거래량, 거래량 보조지표, 추세 보조지표, 채널 보조지표

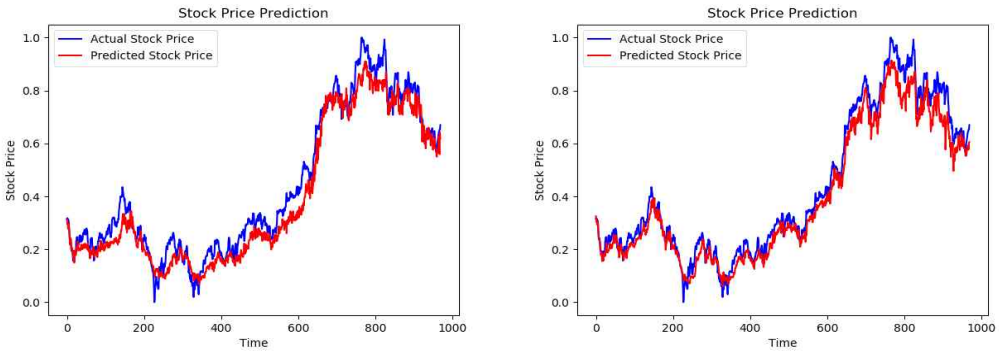

Fig. 3.16 Case 15

16) 가격, 거래량, 가격 보조지표, 거래량 보조지표, 추세 보조지표, 채널 보조지표

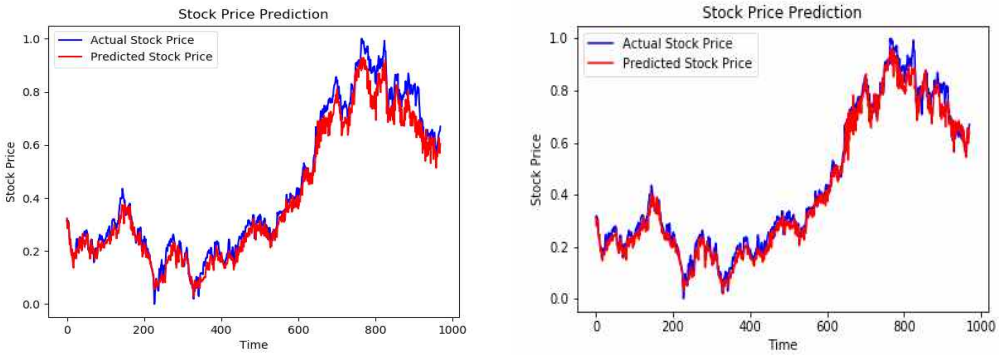

Fig. 3.17 Case 16

3. 성능 검증

본 실험에서는 순환신경망 모델이 예측한 특정한 날의 Kospi 지수 값과 예측하기 전날의 Kospi 지수 종가의 비교를 통해 Kospi 지수의 상승/하락을 예측하고 이를 실제 상승/하락 데이터와의 검증을 통해 예측의 정확도를 검증하였다. Table 3.6은 각 Case들에 대해 산출된 상승/하락 예측의 정확도를 보여준다.

Table 3.6 상승/하락 예측

Case	LSTM	양방향 순환신경망
Case1	50%	50%
Case2	56%	56%
Case3	50%	46%
Case4	50%	49%
Case5	48%	47%
Case6	53%	54%
Case7	59%	59%
Case8	57%	60%
Case9	49%	46%
Case10	51%	54%
Case11	52%	49%
Case12	48%	53%
Case13	48%	56%
Case14	55%	59%
Case15	46%	46%
Case16	48%	50%

상승/하락의 정확도 기준으로 우수한 모델은 Case8, Case7, Case14, Case2 순이었다. 양방향 순환신경망은 LSTM과 비슷한 성능을 내거나 더 우수한 성능을 가지는 것을 볼 수 있었으며 가장 높은 상승/하락의 정확도는 60%이다.

또한 Kospi 지수의 상승/하락 변동이 크게 예측되는 경우에는 상승/하락 예측의 정확도가 더 높아질 수 있을 것이라는 가정 하에 상승/하락률이 각각 0.5%, 1%, 2% 이상이 되는 경우만을 고려하여 정확도를 예측한 결과는 다음 Table 3.8과 Table 3.9에서 보는 바와 같다. Table 3.7은 Table 3.6에서 상승/하락의 예측 정확도가 가장 높은 4개 모델을 나타낸다.

Table 3.7 상승/하락의 예측 정확도가 높은 4개 모델

Case	LSTM	양방향 순환신경망
Case2	56%	56%
Case7	59%	59%
Case8	57%	60%
Case14	55%	59%

Table 3.8 변동이 일정수준 이상인 경우의 LSTM 예측 정확도

	상승/하락	0.5%	1%	2%
Case2	56%	64%	65%	74%
Case7	59%	62%	61%	80%
Case8	57%	64%	69%	77%
Case14	55%	59%	58%	62%

Table 3.9 변동이 일정수준 이상인 경우의 양방향 순환신경망 예측 정확도

	상승/하락	0.5%	1%	2%
Case2	56%	56%	59%	69%
Case7	59%	63%	65%	92%
Case8	60%	59%	60%	74%
Case14	59%	65%	68%	100%

Table 3.8, Table 3.9로부터 상승/하락률이 0.5%, 1%이상, 2% 이상인 경우에 상승/하락 예측의 정확도 기준으로 우수한 모델은 Case14, Case7, Case8, Case2 순임을 볼 수 있다. Case14의 경우 변동이 0.5% 이상일 때 65%, 1% 이상일 때 68%, 2% 이상일 때 100% 정확도를 보였다. Case14와 Case7에 포함된 공통적인 보조지표는 가격 보조지표와 추세 보조지표였다. 이들을 가지고 변동이 일정수준 이상인 경우의 예측 모델을 만든다면 좋은 성능을 낼 것으로 예상된다.

Ⅳ. 결론

다양한 분야에서 순환신경망을 활용한 기술이 확대되고 있다. 자연어 처리, 음성 인식, 주가 예측 등으로 활용하고 있다. 순환신경망은 일반 인공신경망에 시계열 개념을 추가한 것으로, 은닉계층에 이전 정보를 기억시킬 수 있으며 과거의 데이터가 미래 데이터에게 어떻게 영향을 미치는지를 분석한다.

이에 본 논문은 기존 순환신경망을 확장한 LSTM과 양방향 순환신경망을 통해 주가 예측을 실험했다. Table 3.3을 보면 각 보조지표별로 Case를 구성하여 각각의 모델을 평가했다.

각 Case별로 좋은 성능을 낸 Case는 기본(가격+거래량)에 가격 보조지표를 추가한 Case 2이며, Case 8, Case 7, Case 14순이었다. 이들이 갖고 있는 가격 보조지표, 추세 보조지표, 채널 보조지표이다. 모델 선정 시, 가격 보조지표, 추세 보조지표, 채널 보조지표를 쓴다면 모델의 정확성이 높아질 것이다.

본 실험을 통해 얻은 모델이 예측한 값과 예측하기 전날의 종가와 비교를 통해 상승/하락 예측을 검증하였다. 상승/하락 예측의 정확도 기준으로 우수한 모델은 Case8, Case7, Case14, Case2 순이었다. 양방향 순환신경망이 LSTM과 비슷한 성능을 내거나 더 우수한 성능을 가지고 있었다. 가장 높은 상승/하락률은 60%다.

또한 Kospi 지수의 상승/하락 변동이 크게 예측되는 경우에는 상승/하락 예측의 정확도가 더 높아질 수 있을 것이라는 가정 하에 상승/하락률이 각각 0.5%, 1%, 2% 이상이 되는 경우만을 고려하여 정확도를 예측하였다. 상승/하락 예측의 정확도 기준으로 우수한 모델은 Case14, Case7, Case8, Case2 순이었다. Case14의 경우 지수 변동이 0.5% 이상일 때 65%, 1% 이상일 때 68%, 2% 이상일 때 100% 정확도를 보였다. Case14, Case7에 포함된 공통적인 보조지표는 가격 보조지표와 추세 보조지표였다. 이들을 가지고 추세전환에 대한 모델을 만든다면 좋은 성능을 낼 것이다.

향후 연구로는 첫째, 기술적 분석에 한정적이었던 본 논문을 넘어선 기본적 분석과 경제적 지표를 활용한 데이터를 추가한 모델과 본 논문의 모델간의 비교를 할 것이다. 둘째, 추세전환 시뮬레이션에 검증 데이터가 충분하지 않기 때문에 100%라는 정확률이 나왔다. 추후에는 더 많은 데이터와 검증 데이터를 확보하여 트레이딩에 적용할 수 있는 강화학습에 적용할 것이다. 마지막으로 Table 3.2 실험환경에서는 환경적 제약이 있었기 때문에 GPU를 통한 하드웨어 최적화를 통해 알고리즘 성능 개선을 할 것이다.

참고문헌

[1] 주일택. (2018). “양방향 순환신경망을 이용한 주식 가격 예측 모형 연구”, 박사학위논문, 동신대학교 대학원 컴퓨터학과, 전라남도, pp. 12～57.

[2] Khaled A. Althelaya, El-Sayed M. El-Alfy, Salahadin Mohammed (2018). “Evaluation of Bidirectional LSTM for Short- and Long-Term Stock Market Prediction”, Ph.D. Thesis, Department of Information and Computer Science, College of Computer Sciences and Engineering King Fahd University of Petroleum and Minerals, Dhahran 31261, Kingdom of Saudi Arabia, p. 152~156.

[3] 김영숙. (2008). “기본적/기술적 분석을 통한 주식투자전략에 관한 연구”, 석사학위논문, 한양대학교 산업경영디자인대학원 : 경영학과, 서울, pp. 3～25.

[4] 신동하, 최광호, 김창복. (2017). “RNN과 LSTM을 이용한 주가 예측율 향상을 위한 딥러닝 모델”. 한국정보기술학회논문지, 15(10), 9-16.

[5] 김성수, 홍광진. (2017). 순환신경망 기술을 이용한 코스피 200 지수에 대한 예측모델 개발 및 성능 분석 연구. 한국산업정보학회논문지, 22(6), 23-29.

[6] 김호현 (2017). LSTM/GRU 순환신경망을 이용한 시계열데이터 예측. 석사학위논문, 한국방송통신대학교 대학원 : 정보과학과, 서울, 2017. 8, pp. 15~16.

Abstract

A Comparative Study on Stock Price Forecasting Models Using LSTM and Bidirectional Neural Networks

Lee, Jong Hyeok
(Supervisor Kook, Kwang Ho)
Dept. of Software Analysis and Design
Graduate School of
Seoul National University of Science and Technology

Time-series data refers to data having a temporal order such as stocks, sales, and unemployment rate. The forecasting of the time series data is necessary to solve the managerial problems. Through the data forecasting, the performance of the planning and investment can be enhanced, and the cost such as investment opportunity cost, inventory cost, and logistics cost can be reduced.

Stock price prediction is very difficult due to the many variables and the many parties involved. There have been many efforts to predict the stock market. However, there exist limits in linear prediction. Machine learning methods using artificial neural network have been proposed to overcome the limitations of traditional mathematical models. RNN(Recurrent Neural Network) reinforces a time series concept to general neural network and it has been used successfully to analyze time series data. There exist some studies using a bidirectional RNN to predict stock price utilizing price and transaction volume data.

This paper deals with stock price prediction model. We propose to extend RNN method utilizing the auxiliary index of price, transaction volume, trend, and channel in addition to price and transaction volume data. We evaluate the prediction accuracy by comparing the upward/downward prediction of the next days' stock price with the actual upward/downward stock price changes. We

compare the performance of the models using different auxiliary indices and propose the model that yield good performance. Also, assuming that the accuracy of the upward/downward prediction may be higher if the change in stock prices is significantly anticipated, We evaluate models and propose the model that yield good performance when we only consider the cases with upward/downward changes that are more than 0.5%, 1%, and 2% in relative magnitudes, respectively.

감사의 글

본 연구는 서울과학기술대학교 일반대학원 SW분석·설계학과에서 이루어졌으며 이 논문이 이루어지기까지 수많은 사람들의 도움을 받았습니다. 그 분들의 도움덕분에 논문을 쓸 수 있었기에 이 감사의 글에서 감사의 말씀을 드리고 싶습니다.

먼저 논문의 주제 선정부터 실험까지 모든 일을 깊은 애정과 관심으로 지도해주시고 지원을 해주신 국광호 교수님에게 감사의 말씀을 드리고 싶습니다. 그리고 바쁘신 와중에도 학위 논문 심사를 해주신 장성용 교수님, 김우제 교수님께 감사드립니다.

컴퓨터 전공이 아닌 제가 SW분석·설계학과에 들어와서 공부를 할 수 있었던 건 행운이었습니다. 우선 제가 이 학과에 들어올 수 있었던 건 김우제 교수님의 역할이 가장 컸다고 생각합니다. 김우제 교수님께 다시 한 번 감사드립니다. 또한 입사결정을 내려주신 세리정보기술의 오준섭 이사님과 세리정보기술의 제 멘토이신 김자언 이사님, 인턴 때부터 잘 지도해 주신 원 이사님 감사드립니다. 또한 제가 뭘 할지 몰라 해맬 때, 같이 이끌어 준 7기 선배들에게 감사합니다. 특히 7기 실세 건환이, 두 젓가락 라면맨 친구 정은이, 배그 마스터 종희형, 재미 없는 형 경빈이 형, 술 좋아하는 수빈이에게 감사합니다. 세상에 없는 동기들 중 영원한 리더인 친형 원식이형, 이쁘고 날씬하고 귀염둥이 미순, 이쁘고 귀여운 수민, 존지존지존지 조은지 그리고 학과의 행정업무를 맡아준 장예원 조교님께 감사드립니다.

마지막으로 언제나 힘이 되어줬던 엄마, 아빠, Forever My Bro 욱이, Big Bro 누나에게 감사하다고 말씀 드리고 싶습니다. 이제 시작한 신입사원이지만, 20년, 30년 일해서 퀀트 투자가 및 기술사가 되고 싶으며, 후에는 누군가에게 도움이 되는 사람이 되고 싶습니다. 모두에게 감사합니다.