ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 March: 24(1): 167~181 http://dx.doi.org/10.13088/jiis.2018.24.1.167

효과적인 입력변수 패턴 학습을 위한 시계열 그래프 기반 합성곱 신경망 모형: 주식시장 예측에의 응용

이모세
국민대학교 비즈니스IT전문대학원
(*lms0417@kookmin.ac.kr*)

안현철
국민대학교 비즈니스IT전문대학원
(*hcahn@kookmin.ac.kr*)

지난 10여 년간 딥러닝(Deep Learning)은 다양한 기계학습 알고리즘 중에서 많은 주목을 받아 왔다. 특히 이미지를 인식하고 분류하는데 효과적인 알고리즘으로 알려져 있는 합성곱 신경망(Convolutional Neural Network, CNN)은 여러 분야의 분류 및 예측 문제에 널리 응용되고 있다. 본 연구에서는 기계학습 연구에서 가장 어려운 예측 문제 중 하나인 주식시장 예측에 합성곱 신경망을 적용하고자 한다. 구체적으로 본 연구에서는 그래프를 입력값으로 사용하여 주식시장의 방향(상승 또는 하락)을 예측하는 이진분류기로써 합성곱 신경망을 적용하였다. 이는 그래프를 보고 주가지수가 오를 것인지 내릴 것인지에 대해 경향을 예측하는 이른바 기술적 분석가를 모방하는 기계학습 알고리즘을 개발하는 과제라 할 수 있다. 본 연구는 크게 다음의 네 단계로 수행된다. 첫 번째 단계에서는 데이터 세트를 5일 단위로 나눈다. 두 번째 단계에서는 5일 단위로 나눈 데이터에 대하여 그래프를 만든다. 세 번째 단계에서는 이전 단계에서 생성된 그래프를 사용하여 학습용과 검증용 데이터 세트를 나누고 합성곱 신경망 분류기를 학습시킨다. 네 번째 단계에서는 검증용 데이터 세트를 사용하여 다른 분류 모형들과 성과를 비교한다. 제안한 모델의 유효성을 검증하기 위해 2009년 1월부터 2017년 2월까지의 약 8년간의 KOSPI200 데이터 2,026건의 실험 데이터를 사용하였다. 실험 데이터 세트는 CCI, 모멘텀, ROC 등 한국 주식시장에서 사용하는 대표적인 기술지표 12개로 구성되었다. 결과적으로 실험 데이터 세트에 합성곱 신경망 알고리즘을 적용하였을 때 로지스틱회귀모형, 단일계층신경망, SVM과 비교하여 제안모형인 CNN이 통계적으로 유의한 수준의 예측 정확도를 나타냈다.

주제어 : 기술적 분석가, 딥러닝, 분류기, 주가지수 등락 예측, 합성곱 신경망

논문접수일 : 2018년 1월 15일 논문수정일 : 2018년 3월 17일 게재확정일 : 2018년 3월 20일
원고유형 : 일반논문(급행) 교신저자 : 안현철

1. 서론

딥러닝(deep learning)은 사람의 뇌와 유사한 동작방식을 가지고 있는 알고리즘으로써 지난 10여 년간 다양한 기계학습 알고리즘 중에서 많은 주목을 받아왔다. 특히 딥러닝 알고리즘의 일종인 합성곱 신경망(convolutional neural network, CNN)은 이미지 인식 및 분류에 효과적인 성능을 보이고 있다(LeCun et al., 2015). 이러한 딥러닝에 관한 연구는 신경망의 층이 깊어지면서 생기는 막대한 연산량을 GPU(graphic processing unit)를 활용하여 학습시킬 수 있게 되어 기존의 CPU(central processing unit)를 통한 연산 대비 약 10배 이상 빠른 학습이 가능해져 더 활발해지고 있다(Kwon and Kang, 2017).

주가예측은 주식시장에서 발생하는 다양한 요

인들로 인해 예측하기 어려운 문제 중 하나이다. 현업에서는 이른바 기술적 분석 전문가들이 과거의 주가지수나 거래량 등 주식시장에서 나타나는 데이터와 차트를 분석하여 주가동향과 투자심리 등을 예측하려고 시도하고 있다. 학계에서는 인공신경망, SVM(support vector machine) 등 전통적인 기계학습 알고리즘을 이용하여 주식시장에서 기술적 분석을 위해 사용하는 기술지표들을 입력변수로 사용하여 주식시장을 예측하려는 연구들이 주로 발표되어 왔다(Kim et al., 2004; Kim, 2012; Park et al., 2016). 최근에는 뉴스 기사, SNS(social networking service) 등에서 발생하는 텍스트 데이터 분석을 통해 주가지수의 방향성을 예측하려는 연구들(Hong et al., 2016; Jeon et al., 2016; Lee and Lee, 2017)이 발표되고 있으며, 기술지표를 바탕으로 차트를 그려서 주가의 패턴을 찾으려고 시도한 연구도 있었다(Ryu et al., 2017).

본 연구에서는 Ryu et al.(2017)과 같이 기술지표를 시각화하여 주가의 패턴을 찾는 시도를 하려고 한다. 다만 기술지표를 이용하여 방사형차트를 그리고 클러스터링 방법론을 사용한 Ryu et al.(2017)의 연구와 달리, 기술지표들을 입력변수로 하여 꺾은선 그래프를 그려서 만들어진 이미지 데이터를 입력데이터로 한다는 점에서 차별화된다. 구체적으로 본 연구가 제안하는 알고리즘은 먼저 입력변수들을 5일 단위로 슬라이딩 분할하여 그래프를 만들고, 만들어진 이미지 데이터를 입력변수로 하여 이미지 분류에 효과적인 성능을 보이는 합성곱 신경망에 적용하여 주가지수의 등락을 예측하는 것이다.

본 논문은 다음과 같은 순서로 전개된다. 우선 2장에서는 기계학습을 이용한 주가지수 예측 관련 문헌 연구와 본 연구의 토대가 되는 합성곱 신경망에 대한 기초개념과 그래프 기반 패턴 인식 관련 문헌연구를 살펴본다. 3장에서는 합성곱 신경망을 이용한 주가지수 예측 연구모형을 제안한다. 4장에서는 본 연구의 실증분석을 위한 데이터, 실험설계 그리고 실험결과에 대해 논의한다. 마지막으로 5장에서는 본 연구의 결론과 함께 본 연구가 갖는 의의와 한계점, 향후 연구방향을 제시할 것이다.

2. 이론적 배경

2.1 주가지수 예측

주가지수 예측을 위한 방법으로는 회사의 가치를 분석하여 현재 시장가격과의 차이를 밝혀 향후 가격을 예측하는 기본적 분석기법과 시간변화에 따른 시장가격 변동을 연구하여 패턴을 찾아내고 향후 가격을 예측하는 기술적 분석기법이 있다. 그러나 주식시장에서 발생하는 예측하기 어려운 다양한 변수들로 인해 예측 정확성은 뛰어나지 못하다. 이와 같은 문제를 해결하기 위한 방법으로 기계학습 알고리즘을 통한 연구가 활발하게 진행되어 왔다. 주가지수 예측에 관한 주요 연구들은 다음과 같다.

Kim et al.(2004)은 입력변수로 가격변화를 나타내는 Momentum, ROC와 주가의 전환점을 알려주는 %K, %D, Slow %D, CCI, ROC와 장기간의 추세경향을 알아보는 Linear Slope, LU/LD와 이동평균을 나타내는 MACD, WARS 등 기술적 지표를 활용하여 주가지수의 등락을 예측하고자 하였다. 이 때 분류모형으로는 인공신경망과 SVM을 활용하였다.

Roh et al.(2005)은 주가지수의 변동성을 예측

하기 위하여 금융시계열 모형과 인공신경망 모형을 통합한 통합모형을 제시하였다. 이 연구에서는 금융시계열 모형을 바탕으로 예측하고자 하는 KOSPI200에 대한 통계적인 분석을 통해 인공신경망 모형을 위한 입력변수를 도출함으로써 예측 정확도를 향상시켰다.

Lee et al.(2008)은 22개의 표준기본경제지표로부터 통계적 분석을 통해 12개의 경제지표를 추출하고, 추출된 경제지표들을 예측하고자 하는 예측일에 따라 최량부분적합법을 이용하여 다시 한 번 입력변수들을 선정하였다. 이러한 입력변수를 바탕으로 인공신경망을 통해 주가지수를 예측하고자 하였는데, 제안된 방법이 전통적인 방법보다 우수함을 실증 분석하였다.

Kim et al.(2008)은 유전자 알고리즘(genetic algorithm; GA)을 이용하여 주가지수들의 관계를 찾고, 2005년부터 2007년까지의 실제 주가지수를 가지고 모의투자 시뮬레이션하여 모의투자금액이 230% 증가함을 제시하였다.

Kim and Ahn(2010)은 기존 연구들에서 등락의 기준으로 삼는 단일 임계치가 아닌 이중 임계치를 유전자 알고리즘을 이용하여 최적화하고 SVM을 이용하여 주가지수의 등락을 예측하여 트레이딩 시스템의 매수, 매도, 유지의 신호로 해석하는 지능형 트레이딩 시스템을 제안하였다. 수익률 관점에서 다른 모형들과 비교한 결과 더 우수한 모형임을 확인하였다.

Kim(2012)은 개인 주식 투자자의 의사결정 지원을 해주는 데이터마이닝 도구를 제안하였다. 제안한 데이터마이닝 도구의 주가예측 모형은 기술적 분석을 위해 주식시장에서 사용하는 기술지표들을 입력변수로 하는 인공신경망을 사용하였다.

Park et al.(2016)은 시간에 따른 가격 변화의 정도인 변동성을 나타내는 ATR, 시장의 방향성을 나타내는 DMI, KOSPI 지수의 과거 60일 변동성을 나타내는 Volatility와 같은 주식시장에서 사용되는 기술적 지표 6가지를 입력변수로 하여 SVM, 라쏘 회귀분석, 인공신경망 모형을 사용하여 KOSPI지수에 대한 예측정확도를 비교하는 연구를 하였다.

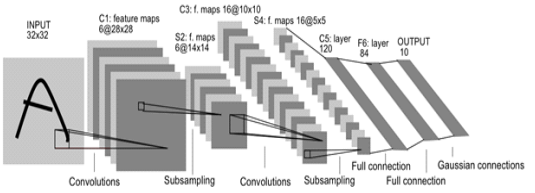
Lee and Ahn(2017)은 유전자 알고리즘을 이용하여 입력변수와 학습사례를 동시에 최적화하고 MSVM(multiclass support vector machine)을 이용하여 주가지수 추세를 상승, 하락, 박스권 세 가지 분류체계로 확장하여 주가지수 추세를 예측하고자 하였는데, 2004년부터 2017년까지의 실제 KOSPI200지수를 이용하여 제안된 방법이 기존의 기계학습 알고리즘보다 성능이 우수함을 확인하였다.

반면, 주가지수의 패턴을 찾기 위한 연구도 시도되었다. Ryu et al.(2017)은 주식시장에 상장된 주식을 대상으로 일별 시가, 저가, 고가, 종가 등을 포함한 총 26가지의 기술적 지표를 바탕으로 방사형 차트를 그려서 주가 패턴을 클러스터화하고, 데이터 시각화를 통하여 각 패턴의 유형을 분석하는 연구를 하였다. 이 때 클러스터링의 방법론은 자기조직화지도(self organizing map, SOM)을 이용하였다.

2.2 합성곱 신경망

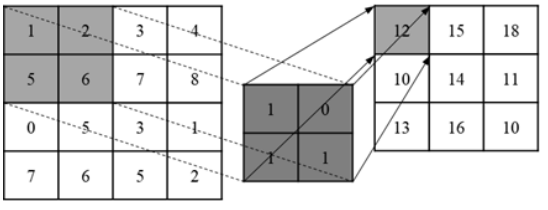
딥러닝 알고리즘의 일종인 합성곱 신경망(convolutional neural network, CNN)은 패턴이나 물체를 인식하는 생물의 시각처리과정을 모방한 모형으로써 LeCun et al.(1998)에 의해 합성곱 신경망이 발전하는 계기가 되었다. 합성곱 신경망은 아래의 <Figure 1>과 같이 하나의 입력계층과

출력계층, 하나 이상의 합성곱 계층(convolution layer)과 풀링 계층(pooling layer) 그리고 완전 연결 계층(fully connected layer)로 구성되어 있다. 입력층을 통해 입력하고자 하는 이미지를 입력하고 합성곱 계층을 통해 필터링되어 적절한 특징을 추출한다.



〈Figure 1〉 Architecture of LeNet-5,
a Convolutional Neural Network
(Adopted from LeCun et al., 1998)

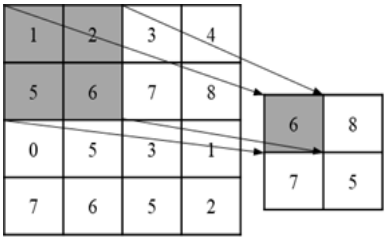
이 때 필터의 개수에 따라 특징맵의 수가 정해진다. 예시로 나타낸 아래의 <Figure 2>와 같이 학습을 통해 변화되는 가중치를 가지고 있는 2×2크기의 합성곱(convolution) 필터가 4×4크기의 입력층에 대해 왼쪽에서 오른쪽으로 위에서 아래로 전체 영역을 훑고 지나가면서, 가중치를 곱하여 합한 결과들이 출력층의 출력값이 된다.



〈Figure 2〉 Convolution filter

합성곱 필터와 함께 사용되는 활성화 함수는 ReLU(rectified linear unit) 활성화 함수로서 음수의 경우는 0으로 양수이면 입력값을 그대로 출력하는 함수이다. 이것은 신경망을 학습시키는 과정에서 출력층에서 멀어질수록 발생하는 신경망의 출력 오차가 반영되지 않는 그래디언트 소실(vanishing gradient)을 해결하기 위한 방안으로 사용된다.

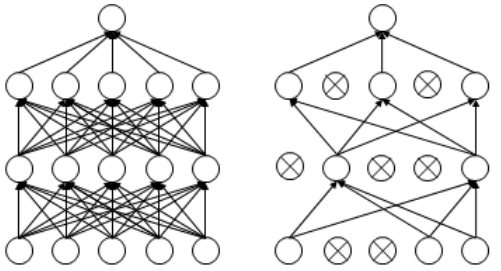
<Figure 1>에서 subsampling은 화면의 크기를 줄이는 과정으로써 합성곱 계층 뒤에 일반적으로 따라오며 출력 값을 단순하게 압축하여 합성곱 계층에서 생산한 정보를 간결(compact)하게 만들어 주고 잡음(noise)을 감소시킬 수 있어 이미지의 분별력을 높일 수 있다. 아래의 <Figure 3>와 같이 맥스풀링(max pooling) 필터에 대해 살펴보면 그림처럼 2×2크기의 필터가 2칸씩 훑고 지나가면서 해당영역에서 최대값을 선택한 결과들이 출력층의 출력값이 된다.



〈Figure 3〉 Max pooling filter

완전 연결 계층에서는 <Figure 4>의 (a)와 같이 완전 연결되어 있는 일반적인 신경망의 은닉계층(hidden layer)을 나타낸다. 은닉 계층에 많은 뉴런을 사용하면 상세한 모형을 만들 수는 있지만 학습용 데이터에 과적합(overfitting)될 수 있다. 이를 예방하기 위해 <Figure 4>의 (b)와 같이 임의로 일정 비율의 뉴런을 제거하면서 학습시키는 드롭아웃(dropout)이라는 기법을 사용한다. 이 드롭아웃 기법은 비슷한 가중치를 가지고 있는 뉴런들이 하나의 뉴런처럼 움직이는 상호적

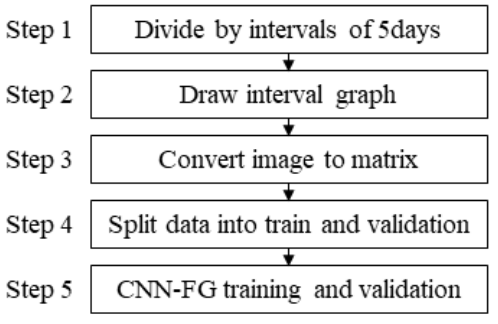
용현상(co-adaptation)을 예방하고 과적합에 효과적이다(Srivastava et al., 2014).



〈Figure 4〉 Dropout neural network model (Adopted from Srivastava et al., 2014)

3. 제안 알고리즘

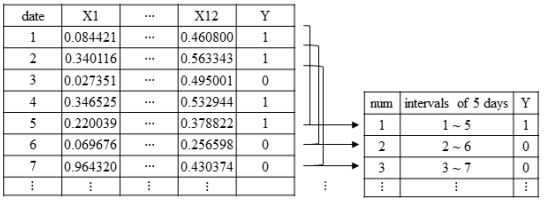
본 연구에서는 독립변수의 일정 간격 그래프 이미지를 입력값으로 반영하여, 전통적인 기계학습 알고리즘보다 개선된 예측 정확도를 보이는 합성곱 신경망 알고리즘을 제안한다. 본 연구에서 제안하는 알고리즘은 편의상 CNN-FG (Convolutional Neural Network-based Fluctuation Graph)로 명명하였다. CNN-FG는 다음의 <Figure 5>에 제시된 것처럼, 총 5단계의 절차로 구현된다.



〈Figure 5〉 Procedure of CNN-FG

1단계: 데이터 세트 5일 간격 슬라이딩 분할

본 연구가 제안하는 CNN-FG의 첫번째 단계는 다음의 <Figure 6>에 제시된 것처럼, 데이터 세트의 독립변수들 X1~X12에 대해서 5일 간격으로 슬라이딩 분할을 한다. 이 때 종속변수는 마지막 5일차에 해당하는 값으로 한다. 이것은 5일 동안의 독립변수의 변동 후 다음 날 주가지수가 오를지 내릴지에 대한 값이다. 5일을 기준으로 데이터세트를 나눈 이유는 주식시장이 장을 여는 날을 기준으로, 1주일 간격으로 그래프를 만들기 위해 설정하였다.



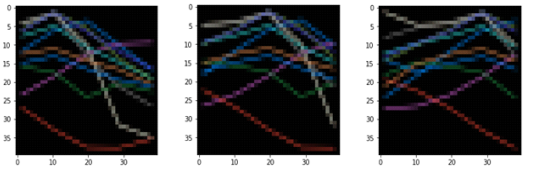
date	X1	···	X12	Y
1	0.084421	···	0.460800	1
2	0.340116	···	0.563343	1
3	0.027351	···	0.495001	0
4	0.346525	···	0.532944	1
5	0.220039	···	0.378822	1
6	0.069676	···	0.256598	0
7	0.964320	···	0.430374	0
⋮	⋮	⋮	⋮	

num	intervals of 5 days	Y
1	1 ~ 5	1
2	2 ~ 6	0
3	3 ~ 7	0
⋮	⋮	⋮

〈Figure 6〉 Division by intervals of 5 days

2단계: 분할된 데이터를 이용한 그래프 생성

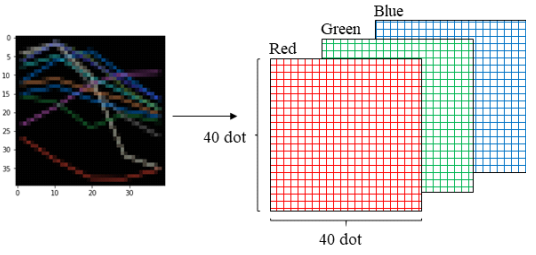
이 단계에서는 1단계에서 5일 간격으로 슬라이딩 분할한 데이터를 가지고 x축은 일 단위, y축은 독립변수들의 값으로 하여 그래프를 그린다. 그래프가 그려진 이미지의 크기는 44, dpi = 10이다. 여기서 dpi는 1에 들어가는 점의 개수를 의미하며 이미지는 가로 40dot, 세로 40dot로 이루어져 있다. 이미지는 RGB색상으로 각 그래프별로 임의의 색이 결정되어 그려졌다. 그래프를 생성하면 아래 <Figure 7>와 같이 그려진다. 아래 그림에서 num=1은 1~5일의 독립변수의 변동 그래프를, num=2는 2~6일의 독립변수의 변동 그래프를, num=3는 3~7일의 독립변수의 변동 그래프를 보여준다.



〈Figure 7〉 Draw interval graph

3단계: 그래프 이미지를 행렬로 변경

이 단계에서는 <Figure 8>과 같이 이미지를 RGB색상에 따라 행렬로 변환한다. 이 작업을 통해 이미지는 40×40×3크기의 행렬로 변환된다. 일반적으로 이미지에 나타나는 색깔은 0~255의 크기를 가지고 있는데 0~1 사이로 정규화(min-max normalization)하여 행렬로 변환한다.



〈Figure 8〉 Convert image to matrix

4단계: 학습용, 검증용 데이터 세트

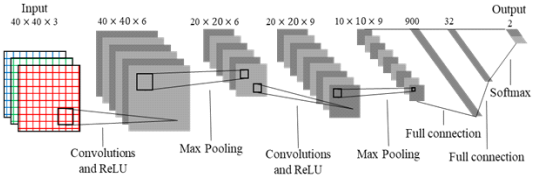
이 단계에서는 3단계에서 만들어진 이미지별 행렬 데이터들을 입력 데이터 세트로 하여, CNN-FG의 일반화 정도를 측정하기 위하여 학습용과 검증용 데이터 세트로 구분한다. 데이터 세트의 구성은 8:2의 비율로 하여 전체의 80%를 차지하는 데이터를 학습용으로 사용하였고, 전체의 20%를 검증용으로 사용하였다.

5단계: CNN-FG 학습 및 검증 예측정확도 도출

5단계에서는 CNN-FG의 신경망을 구성하기 위해서 텐서플로 공식사이트(TensorFlow, https://www.tensorflow.org/versions/master/get_started/mnist/pros)에 제공되어 있는 MNIST(Mixed national institute of standards and technology) 손 글씨 분석을 위한 예제 코드를 활용하였다(Torres, 2017). 아래의 <Table 1>을 살펴보면 CNN-FG를 구성하는 계층의 종류와 그에 따른 크기와 활성화함수 등 설계 변수들에 대한 설정값을 확인할 수 있고, <Figure 9>는 CNN-FG 신경망의 전반적인 구축 과정을 보여준다.

〈Table 1〉 Parameters of CNN-FG

Layer	Size	Activation function
Input layer	N×40×40×3	-
Convolution layer	5×5×6 Convolution filter 5×5×9 Convolution filter	ReLU
Pooling layer	2×2 Max pooling filter	-
Hidden layer	900 Nodes (first layer) 32 Nodes (second layer)	ReLU
Output layer	2 Nodes	Softmax



〈Figure 9〉 Building Process of CNN-FG

4. 실험설계 및 실증분석

4.1 실험 데이터

본 연구에서 제안하는 CNN-FG의 유용성을 검증하기 위해 구글 파이낸스(Google Finance , https://www.google.com/finance)에서 제공하는 일별 KOSPI200 주가 지수를 사용하였다. 모형 구축을 위해 사용한 표본 데이터는 2009년 1월부터 2017년 2월까지 약 8년간 발생한 2,026건의 KOSPI200의 일별 주가지수이다. 지난 2007년 미국에서 시작된 서브프라임 모기지 사태(subprime mortgage crisis)는 국제 금융 시장에 대규모의 금융위기를 야기하였기 때문에 2007~2008년의 주식시장 데이터는 2007년 이전과 2008년 이후와는 상당히 다른 패턴을 보이는 이상치에 가깝다. 따라서 본 연구에서는 글로벌 위기가 해소되기 시작한 2009년부터 시작되는 데이터를 사용하였다(Han, 2017).

본 연구에서 사용된 입력변수는 기술적 분석 전문가들의 관련 연구 검토를 통해 <Table 2>과 같이 주식시장에서 사용하는 12가지 기술지표를 선택하였다(Ahn and Kim, 2008).

종속변수는 종가 지수의 일일 변화 방향에 따라 분류하였다. '0'은 다음날의 종가 지수가 오늘의 지수보다 낮고 '1'은 다음날의 지수가 오늘의 지수보다 높다는 것을 의미한다.

본 연구에서 제안한 CNN-FG의 알고리즘을 살펴보면 1주일 단위로 데이터 세트를 묶기 위하여 주식시장이 1주일 중 개장하는 평일 5일을 기준으로 하여, <Figure 6>와 같이 5일 단위로 데이터 세트를 슬라이딩 분할을 하였다. 따라서 총 2,026건의 데이터 세트를 통해 만들어진 그래프 이미지 데이터의 개수는 총 2,022건이 된다.

〈Table 2〉 Attributes and their formula

Names of feature	Formula
Stochastic %K	$\frac{C_t - LL_{t-5}}{HH_{t-5} - LL_{t-5}} \times 100$
Stochastic %D	$\frac{\sum_{i=0}^{n-1} \%K_{t-i}}{n}$
Stochastic slow %D	$\frac{\sum_{i=0}^{n-1} \%D_{t-i}}{n}$
Momentum	$C_t - C_{t-4}$
ROC (rate of change)	$\frac{C_t}{C_{t-n}} \times 100$
LW %R (Larry William's %R)	$\frac{H_n - C_t}{H_n - L_n} \times 100$
A/D oscillator (accumulation/distribution oscillator)	$\frac{H_t - C_t - 1}{H_n - L_t}$
Disparity 5 days	$\frac{C_t}{MA_5} \times 100$
Disparity 10 days	$\frac{C_t}{MA_{10}} \times 100$
OSCP (price oscillator)	$\frac{MA_5 - MA_{10}}{MA_5}$
CCI (commodity channel index)	$\frac{M_t - SM_t}{0.015 \times D_t}$
RSI (relative strength index)	$100 - 100/(1 + \frac{\sum_{i=0}^{n-1} Up_{t-i}/n}{\sum_{i=0}^{n-1} Dw_{t-i}/n})$

C: closing price; L: low price; H: high price;
LL_n: lowest low price in the last n days;
HH_n: highest high price in the last n days;
M : moving average of price;

$$M_t = \frac{H_t + L_t + C_t}{3};\ SM_t = \frac{\sum_{i=0}^{n} M_{t-i+1}}{n};$$

$$D_t = \frac{\sum_{i=1}^{n} |M_{t-i+1} - SM_t|}{n};$$

Up: upward price change; Dw: downward price change

종속변수는 종가 지수의 일일 변화 방향에 따라 분류하였다. ‘0’은 다음날의 종가 지수가 오늘의 지수보다 낮고 ‘1’은 다음날의 지수가 오늘의 지수보다 높다는 것을 의미한다.

본 연구에서 제안한 CNN-FG의 알고리즘을 살펴보면 1주일 단위로 데이터 세트를 묶기 위하여 주식시장이 1주일 중 개장하는 평일 5일을 기준으로 하여, <Figure 6>와 같이 5일 단위로 데이터 세트를 슬라이딩 분할을 하였다. 따라서 총 2,026건의 데이터 세트를 통해 만들어진 그래프 이미지 데이터의 개수는 총 2,022건이 된다. 일반화된 연구 모형을 만들기 위해 종속변수 ‘1’과 ‘0’의 비율을 1:1로 맞추기 위해 데이터를 살펴본 결과 데이터는 ‘1’과 ‘0’의 개수는 각각 1,047개, 975개로 나타났다. 따라서 1,047개의 중에서 랜덤샘플링(random sampling)하여 ‘1’과 ‘0’의 비율을 1:1로 만들었다. ‘1’과 ‘0’의 개수가 각각 975개씩 되어 학습용과 검증용 데이터 세트로 나누기 위해 8:2의 비율로 나누고 합친 결과 아래 <Table 3>와 같이 학습용 데이터 세트는 1560개, 검증용 데이터 세트는 380개가 되었다.

〈Table 3〉 Training and validation dataset

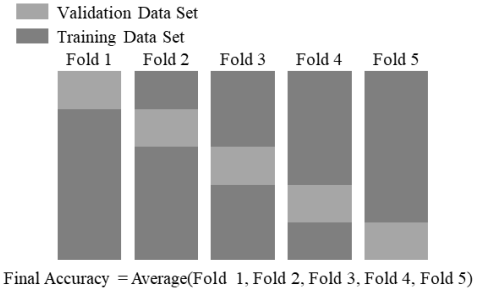
	Up(‘1’)	Down(‘0’)	Total
Training	780	780	1560
Validation	195	195	390
Total	975	975	1950

4.2 실험설계

CNN-FG의 신경망을 구성하기 위해 텐서플로우(TensorFlow)를 활용하여 알고리즘을 구현하였다. 텐서플로우는 2015년 구글 브레인 팀에서 오픈 소스로 공개한 머신 러닝 시스템으로써 파이썬(python) 라이브러리의 형태로 문법이 개발되어 있는 딥러닝 프레임워크이다(Chung et al., 2017).

제안 알고리즘의 성능을 면밀하게 확인하기 위해, 본 연구에서는 전통적인 기계학습 알고리즘과 비교하였다. 비교모형으로 선정한 모형은 로지스틱 회귀모형(logistic regression), SVM (support vector machine), 인공신경망(artificial neural network; ANN)이다.

기계학습을 할 때 사용한 데이터의 양이 충분치 않아 분류기 성능측정의 통계적 신뢰도를 높이기 위해 쓰는 방법 중 k겹 교차검증(k-fold cross validation)을 사용하였다. 본 연구에서는 아래의 <Figure 10>와 같이 데이터 세트를 5개의 집단으로 나누어 4개의 집단으로 학습하고 1개의 집단으로 검증을 하는 5겹 교차검증을 사용하였다.



〈Figure 10〉 5-fold cross validation

4.3 실험결과

앞 절에서 설명한 실험설계를 통해 모형별 주가지수 등락예측 실험을 수행하였다. <Table 4>에는 실험한 모형들의 각 Fold별 가장 높은 결과를 보인 모형의 설정 값에 대한 표이다. CNN-FG

의 경우에는 100 step 단위로 출력되는 예측정확도 중에서 가장 높은 정확도를 채택하였다. 또한 드롭아웃의 경우 노드가 드롭아웃되지 않을 확률 70%를 기본 설정으로 하였다.

〈Table 4〉 Training and validation dataset

Dataset	LOGIT	ANN	SVM	CNN-FG
Fold 1	Backward elimination	hidden layer = 24	RBF kernel, C=1, =1	Step=1100
Fold 2	Backward elimination	hidden layer = 24	Polynomial kernel, C=78, =4	Step=800
Fold 3	Backward elimination	hidden layer = 24	Polynomial kernel, C=78, =2	Step=900
Fold 4	Backward elimination	hidden layer = 12	Polynomial kernel, C=78, =5	Step=1000
Fold 5	Enter method	hidden layer = 18	linear kernel, C=55	Step=400

〈Table 5〉 Results

Dataset		LOGIT	ANN	SVM	CNN-FG
Fold 1	Train	52.88%	52.24%	54.17%	55.26%
	Valid	54.10%	52.05%	57.44%	58.72%
Fold 2	Train	54.04%	50.51%	52.50%	58.97%
	Valid	49.74%	51.03%	51.03%	57.69%
Fold 3	Train	52.12%	52.56%	53.21%	55.83%
	Valid	53.85%	55.13%	56.41%	57.44%
Fold 4	Train	54.36%	50.90%	52.56%	55.38%
	Valid	51.28%	53.33%	53.59%	57.44%
Fold 5	Train	54.29%	51.54%	54.10%	54.74%
	Valid	52.82%	52.82%	54.62%	57.44%

<Table 5>는 제안 모형인 CNN-FG과 로지스틱 회귀분석, 인공신경망, SVM의 총 4가지의 비교모형 실험결과를 종합한 표이다. 실험 결과에 따르면 제안 모형의 예측 정확도가 평균 57.74%로 기존 기법들에 비해 개선된 예측 정확도를 보이는 것을 확인하였다. 또한 학습용 데이터 세트와 검증용 데이터 세트에 대한 모형 간의 예측 정확도 차이가 적어서 상당히 안정된 결과를 보여주었다. 이를 통해 그래프 이미지를 입력 데이터로 하는 합성곱 신경망을 이용한 주가지수 등락 예측모형구축은 효과적임을 증명하였다.

이어서 제안 모형과 비교 모형들 간의 예측 정확도의 차이가 통계적으로 유의한지를 확인하기 위해 Two-Sample Test for Proportions를 실시하였다. 그 결과 아래의 <Table 6>와 같이 본 연구에서 제안한 CNN-FG는 로지스틱 회귀분석과 99%, 인공신경망과 95%, SVM과 95% 신뢰수준 하에서 통계적으로 유의한 차이를 보이고 있다. 따라서 제안 모형과 비교 모형들 간의 예측 정확도의 차이가 통계적으로 유의함을 확인할 수 있다.

〈Table 6〉 Training and validation dataset

	ANN	SVM	CNN-FG
LOGIT	-0.321	-1.413*	-3.380***
ANN		-1.092	-3.060**
SVM			-1.969**

* statistical significant at 10%, ** statistical significant at 5%, *** statistical significant at 1%

5. 결론

본 연구에서는 예측하기 어려운 주가지수 등락 문제를 해결하기 위해서 그래프의 변동을 보고 주가지수가 오를 것인지 내릴 것인지에 대해 경향을 예측하는 이른바 기술적 분석가를 모방하는 기계학습 알고리즘을 개발하는 모형을 제안하였다. CNN-FG로 명명된 본 연구의 제안모형은 이미지의 인식과 분류에 뛰어나다고 알려져 있는 딥러닝의 알고리즘의 일종인 합성곱 신경망을 이용하여 주식시장에서 사용하는 주요 기술지표들의 5일 간의 변화 양상을 그래프로 만들어 그 그래프를 입력 데이터로 사용하고자 하였다. 제안모형의 주가예측 적용 가능성을 확인하고자 실제 데이터에 로지스틱 회귀분석, 인공신경망, SVM과 같은 비교모형과 제안모형을 동시에 적용해 본 결과, 제안모형이 가장 우수한 예측 정확도를 보여주었고, 그 차이는 통계적으로 유의한 것으로 확인하였다.

본 연구가 가지는 의의는 학술적 관점에서 국내 주식시장에서 CNN을 이용한 첫 시도라는 점이다. 또한 그래프를 입력데이터로 사용하여 기술적 분석가를 모방하는 기계학습 알고리즘을 제안했다는 점에서 의의를 갖는다. 실무적으로는 시계열에 따라 변하는 다른 금융상품에 대해서도 적용해 볼 수 있는 새로운 해법을 제시하였다는 점에서 의의를 갖는다.

그러나 본 연구는 다음과 같은 몇가지 한계점을 갖는다. 먼저 본 연구의 제안모형은 합성곱 신경망에 대한 기본적인 알고리즘을 가지고 실험을 함으로써 다양한 조건에서 실험해보지 못한 설계 상의 한계점을 가지고 있다. 그러므로 향후 연구에서는 합성곱 신경망 알고리즘을 다양한 조건에서 실험해보고 입력 데이터에 따른 최적의 설계 조건을 찾아보는 연구가 필요하다.

두번째로는 제안모형은 5일치의 데이터가 한 개의 데이터로 입력되었는데 다른 비교모형들은 동일한 조건으로 실험한 것이 아니므로 엄밀히 말하면 공정하지 않은 부분이 있다. 향후 연구에서는 시계열에 대한 분석에 효과적인 딥러닝의 일종인 순환 신경망(recurrent neural network; RNN)과 같은 기계학습 알고리즘과의 정밀한 비교모형실험이 필요하다.

세번째로 독립변수 12가지에 대해서 한 이미지에 12가지의 그래프를 다 그려서 실험을 진행하였는데 각각의 독립변수의 변동 그래프가 예측 정확도에 영향을 얼마나 주는지에 대한 정밀한 검증이 없었다. 따라서 향후 연구에서는 영향력이 있는 독립변수의 그래프를 이용하여 주가지수의 등락을 예측하는 연구가 필요하다. 또한 주가지수에 영향을 주는 독립변수는 외부에서 발생되는 변수들도 존재하는데 내부에서 파생된 변수만을 가지고 실험을 진행하였다. 따라서 향후 연구에서는 환율, 물가지수 등 다양한 주식시장 외부변수들과 함께 실험을 진행할 필요가 있다.

마지막으로 본 연구의 제안모형은 주가지수 등락예측이라는 특정분야에 대해서만 실험이 진행되었기 때문에 그래프의 변동성을 가지고 예측을 하는 알고리즘의 일반성이 충분히 검증되지 못하였다. 따라서 사람의 감정변화에 대한 그래프나 기존에 다른 시계열 비즈니스 문제에서 사용한 독립변수들에 대한 그래프와 같은 다양한 분야에서 제안모형의 일반성을 검증하는 것이 향후 연구과제가 될 수 있을 것이다.

참고문헌(References)

Ahn, H. and K.-j. Kim, "Using genetic algorithms to optimize nearest neighbors for data mining," Annals of Operations Research, Vol.163, No.1(2008), 5-18.

Chung, Y., S.M. Ahn., J. Yang, and J. Lee, "Comparison of Deep Learning Frameworks: About Theano, Tensorflow, and Cognitive Toolkit," Journal of Intelligence and Information Systems, Vol.23, No.2(2017), 1-17.

Han, H. W., "An Intelligent System Trading using Optimized CBR with Absolute Similarity Threshold," Ph.D. dissertation, The Graduate School of Business IT, Kookmin University, 2017

Hong, T., T. Lee, and J. Li, "Development of Sentiment Analysis Model for the hot topic detection of online stock forums," Journal of Intelligence and Information Systems, Vol.22, No.1(2016), 187~204.

Jeon, S., Y. Chung, and D. Lee, "The Relationship between Internet Search Volumes and Stock Price Changes," Journal of Intelligence and Information Systems, Vol.22, No.2(2016), 81~96.

Kim, S., D. Kim, C. Han, and W. Kim, "Stock Forecasting using Stock Index Relation and Genetic Algorithm," Journal of Korean Institute of Intelligent Systems, Vol 18, No.6(2008), 781-786.

Kim, S.-D., "Data Mining Tool for Stock Investors' Decision Support," Journal of The Korea Contents Association, Vol.12, No.2(2012), 472-482.

Kim, S.-W., and H. Ahn, "Development of an Intelligent Trading System Using Support Vector Machines and Genetic Algorithms," Vol.16, No.1(2010), 71-92.

Kim, Y., E. Shin, and T. Hong, "Comparison of Stock Price Index Prediction Performance Using Neural Networks and Support Vector Machine," The Journal of Internet Electronic Commerce Research, Vol.4, No.3(2004), 221-243.

Kwon, D.-C., and B.-Y. Kang, "CPU and GPU Performance Analysis for Convolution Neural Network," The Journal of Korean Institute of Information Technology, Vol.15, No.8(2017), 11-18.

LeCun, Y., L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the Institute of Electrical and Electronics Engineers, Vol.86, No.11(1998), 2278~2324.

LeCun, Y., Y. Bengio, and G. Hinton, "Deep learning," Nature, Vol.521, No.7553(2015), 436-444.

Lee, E. J., C. H. Min, and T. S. Kim, "Development of the KOSPI (Korea Composite Stock Price Index) forecast model using neural network and statistical methods," The Institute of Electronics Engineers of Korea - Computer and Information, Vol.45, No.5(2008), 95-101.

Lee, J.-D., and H. Ahn, "A Study on the Prediction Model of Stock Price Index Trend based on GA-MSVM that Simultaneously Optimizes Feature and Instance Selection," Journal of Intelligence and Information Systems, Vol.23, No.4(2017), 147-168.

Lee, M.-S., and H. J. Lee, "Stock Price Prediction by Utilizing Category Neutral Terms : Text

Mining Approach," Journal of Intelligence and Information Systems, Vol.23, No.2 (2017), 123-138.

Park, J. Y., J. R, and H. J. Shin, "Predicting KOSPI Stock Index using Machine Learning Algorithms with Technical Indicators," Journal of Information Technology and Architecture, Vol.13, No.2(2016), 331-340.

Roh, T. H., T. H. Lee, and I. G. Han, "Forecasting the volatility of KOSPI 200 Using Neural Network-financial Time Series Model," Korean Management Review, Vol.34, No.3 (2005), 683-713.

Ryu, J. P., H. J. Shin, M. H. Kim, and J. Baek, "Pattern Analysis of Stock Prices Using Machine Learning and Data Visualization," Journal of Information Technology and Architecture, Vol.14, No.2(2017), 189-197.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research, Vol.15, No.2(2014), 1929-1958.

Torres, J., "First Contact with TensorFlow", Hanbit Media Inc., 2017.

Abstract

A Time Series Graph based Convolutional Neural Network Model for Effective Input Variable Pattern Learning : Application to the Prediction of Stock Market

Over the past decade, deep learning has been in spotlight among various machine learning algorithms. In particular, CNN(Convolutional Neural Network), which is known as the effective solution for recognizing and classifying images or voices, has been popularly applied to classification and prediction problems. In this study, we investigate the way to apply CNN in business problem solving. Specifically, this study propose to apply CNN to stock market prediction, one of the most challenging tasks in the machine learning research. As mentioned, CNN has strength in interpreting images. Thus, the model proposed in this study adopts CNN as the binary classifier that predicts stock market direction (upward or downward) by using time series graphs as its inputs. That is, our proposal is to build a machine learning algorithm that mimics an experts called 'technical analysts' who examine the graph of past price movement, and predict future financial price movements.

Our proposed model named 'CNN-FG(Convolutional Neural Network using Fluctuation Graph)' consists of five steps. In the first step, it divides the dataset into the intervals of 5 days. And then, it creates time series graphs for the divided dataset in step 2. The size of the image in which the graph is drawn is 40 (pixels) × 40 (pixels), and the graph of each independent variable was drawn using different colors. In step 3, the model converts the images into the matrices. Each image is converted into the combination of three matrices in order to express the value of the color using R(red), G(green), and B(blue) scale. In the next step, it splits the dataset of the graph images into training and validation datasets. We used 80% of the total dataset as the training dataset, and the remaining 20% as the validation dataset. And then, CNN classifiers are trained using the images of training dataset in the final step. Regarding the parameters of CNN-FG, we adopted two convolution filters (5 × 5 × 6 and 5 × 5 × 9) in the convolution layer. In the

* Master's Candidate, Graduate School of Business IT, Kookmin University
** Corresponding Author: Hyunchul Ahn
Graduate School of Business IT, Kookmin University
77, Jeongneung-ro, Seoungbuk-gu, Seoul 02707, Republic of Korea
Tel: +82-2-910-4577, Fax: +82-2-910-4017, E-mail: hcahn@kookmin.ac.kr

Mo-Se Lee* · Hyunchul Ahn**

pooling layer, 2 × 2 max pooling filter was used. The numbers of the nodes in two hidden layers were set to, respectively, 900 and 32, and the number of the nodes in the output layer was set to 2(one is for the prediction of upward trend, and the other one is for downward trend). Activation functions for the convolution layer and the hidden layer were set to ReLU(Rectified Linear Unit), and one for the output layer set to Softmax function.

To validate our model - CNN-FG, we applied it to the prediction of KOSPI200 for 2,026 days in eight years (from 2009 to 2016). To match the proportions of the two groups in the independent variable (i.e. tomorrow's stock market movement), we selected 1,950 samples by applying random sampling. Finally, we built the training dataset using 80% of the total dataset (1,560 samples), and the validation dataset using 20% (390 samples). The dependent variables of the experimental dataset included twelve technical indicators popularly been used in the previous studies. They include Stochastic %K, Stochastic %D, Momentum, ROC(rate of change), LW %R(Larry William's %R), A/D oscillator(accumulation/distribution oscillator), OSCP(price oscillator), CCI(commodity channel index), and so on. To confirm the superiority of CNN-FG, we compared its prediction accuracy with the ones of other classification models. Experimental results showed that CNN-FG outperforms LOGIT(logistic regression), ANN(artificial neural network), and SVM(support vector machine) with the statistical significance. These empirical results imply that converting time series business data into graphs and building CNN-based classification models using these graphs can be effective from the perspective of prediction accuracy. Thus, this paper sheds a light on how to apply deep learning techniques to the domain of business problem solving.

Key Words : Classifier, Convolutional Neural Network, Deep Learning, Stock Price Fluctuation Prediction, Technical Analyst

Received : January 15, 2018 Revised : March 17, 2018 Accepted : March 20, 2018
Publication Type : Regular Paper(Fast-track) Corresponding Author : Hyunchul Ahn

저 자 소 개

이 모 세

국민대학교 기계시스템공학부에서 학사학위를 취득하였으며, 현재 국민대학교 비즈니스IT전문대학원에서 비즈니스IT전공으로 석사과정에 재학 중이다. 주요 관심분야는 비즈니스 애널리틱스, 인공지능 등이다.

안 현 철

현재 국민대학교 경영대학 경영정보학부 부교수로 재직 중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심분야는 금융 및 고객관계관리 분야의 인공지능 응용, 정보시스템 수용과 관련한 행동 모형 등이다.