Advanced Industrial SCIence
Vol. 3, No. 1, pp. 17-22, 2024
ISSN 2951-2476
https://doi.org/10.23153/AI-Science.2024.3.1.017

시계열 데이터 활용에 관한 동향 연구

최신형
강원대학교 전기제어계측공학부 교수

A Study on Trend Using Time Series Data

Shin-Hyeong Choi
Professor, Division of Electrical, Control & Instrumentation, Kangwon National University

요 약 인류의 출현과 함께 시작된 역사에는 기록이라는 수단이 있기에 현재에 사는 우리는 데이터를 통해 과거를 확인할 수 있다. 생성되는 데이터는 일정 순간에만 발생하여 저장될 수도 있지만, 과거로부터 현재까지 일정 시간 간격 동안 계속해서 생성될 뿐만 아니라 다가올 미래에도 발생함으로써 이를 활용하여 예측하는 것 또한 중요한 작업이다. 본 논문은 수많은 데이터 중에서 시계열 데이터의 활용 동향을 알아보기 위해서 시계열 데이터의 개념에서부터 머신러닝 분야에서 시계열 데이터 분석에 주로 사용되는 Recurrent Neural Network와 Long-Short Term Memory에 대해 분석하고, 이런 모델들을 활용한 사례의 조사를 통해 의료 진단, 주식 시세 분석, 기후 예측 등 다양한 분야에 활용되어 높은 예측 결과를 보이고 있음을 확인하였고, 이를 바탕으로 향후 활용방안에 대하여 모색해본다.

주제어 : 데이터, 시계열 데이터, 머신러닝, 순환신경망, LSTM

Abstract History, which began with the emergence of mankind, has a means of recording. Today, we can check the past through data. Generated data may only be generated and stored at a certain moment, but it is not only continuously generated over a certain time interval from the past to the present, but also occurs in the future, so making predictions using it is an important task. In order to find out trends in the use of time series data among numerous data, this paper analyzes the concept of time series data, analyzes Recurrent Neural Network and Long-Short Term Memory, which are mainly used for time series data analysis in the machine learning field, and analyzes the use of these models. Through case studies, it was confirmed that it is being used in various fields such as medical diagnosis, stock price analysis, and climate prediction, and is showing high predictive results. Based on this, we will explore ways to utilize it in the future.

Key Words : Data, Time Series Data, Machine Learning, Recurrent Neural Network, LSTM

1. 서론

우리는 데이터를 통해 과거를 확인할 수 있으므로, 데이터는 과거로부터 축적되어 온 인류의 발자취라고도 할 수 있다. 최근의 4차 산업혁명에서의 핵심 키워드 중에 하나인 빅데이터(Big data)에서도 볼 수 있듯이 이런 데이터는 세월 또는 시간이 흐를수록 쌓이는 양은 엄청나며, 우리가 살고 있는 현재에도 계속해서, 앞으로 다가올 미래에도 끊임없이 생성될 것이다. 이런 데이터를 좀 더 세분하여 살펴본다면, 데이터(Data)와 정보(Information)로 구분할 수 있다. 여기서 데이터란 우리가 살고 있는 세계에서 수집된 사실이나 값으로

*Corresponding Author : Shin-Hyeong Choi(cshinh@kangwon.ac.kr)
Received December 10, 2024
Revised January 15, 2024
Accepted March 20, 2024
Published March 30, 2024

정의할 수 있고, 정보란 이런 데이터를 가공하여 특정한 목적에 맞게 사용할 수 있도록 만든 것이다[1]. 예를 들어서 설명하자면, 데이터는 검색엔진인 네이버, 다음, 구글 등에서 수집하여 저장된 것이라고 볼 수 있으며, 그에 반해 정보는 맛집 리스트와 같이 이들 검색엔진에 저장된 데이터 중에서 필요한 것만 필터링되어 보여지는 것이라고 할 수 있다. 결론적으로, 우리 일상생활의 수많은 데이터 중에서 현재 나의 관심 분야나 필요한 정보로 이루어진다. 또한, 과거의 문헌이나 책자를 통해 정리되던 데이터는 컴퓨터의 발명과 이런 데이터를 저장하는 기술의 발달로 인해 디지털화되었고, 과거의 문헌 데이터 또한 디지털로 변환되어 축적되고 있으므로, 데이터를 지칭할 때는 현재와 미래의 데이터뿐만 아니라 과거의 데이터를 포함하므로 인력의 역사를 생각한다면 그 규모를 수치화하기는 쉽지 않은 작업이다. 빅토어 마이어 쇤베르거(Viktor Mayer-Schonberger)와 케네스 쿠키어(Kenneth Cukier)가 저술한 “빅 데이터가 만드는 세상”에서 “데이터가 폭발하는 거대한 변화의 소용돌이 속에서 우리는 어떻게 살아갈 것인가?”라는 물음에서 알 수 있듯이 엄청난 양의 데이터가 구축됨으로써 우리의 삶이 급속도로 바뀌고 있음을 실감할 수 있는 세상이다[2]. 데이터 구축의 주체 또한 정부, 기관 등의 단체나 기업에서부터 개인이 생성하는 범위까지 다양한 곳으로부터 수집되고 있으므로 이와 더불어 데이터의 종류 또한 다양화되고 정형화된 데이터뿐만 아니라 비정형화된 데이터 양도 많은 비율을 차지하고 있다. 앞서 언급하였듯이 생성되는 데이터는 일정 순간에만 발생하여 저장될 수도 있지만, 과거로부터 현재까지 일정 시간 간격 동안 계속해서 생성될 뿐만 아니라 다가올 미래에도 발생함으로써 이를 예측하는 것 또한 중요한 작업이다. 이와 같이 일정 기간에 대해 시간의 함수로 표현되는 데이터를 시계열 데이터라고 부르며, 경제, 기상, 금융 등 다양한 분야에서 생성되어, 시계열 분석과 예측을 통해 과거 동향을 이해하고 미래 값을 예측하여 의사 결정을 지원하는 데 활용되고 있다.

본 논문은 수많은 데이터 중에서 시간정보를 저장하고 있는 시계열 데이터(Time series data)의 활용 동향을 살피고 향후 이를 활용하기 위한 시사점을 도출하고자 하며, 논문의 구성은 다음과 같다. 2장에서는 데이터의 개념과 함께 빅데이터와 시계열 데이터 등에 대해서 설명하고, 3장에서는 이를 기반으로 활용되고 있는 기술 과 서비스에 대해서 기술하며, 인공지능 기술을 활용한 시계열 데이터 분석 사례를 제시하고, 마지막은 결론으로 정리한다.

2. 이론적 고찰

2.1 데이터

서론에서도 기술하였듯이, 현대사회는 정보사회라고 부를 만큼 다양한 분야나 조직과 더불어 수많은 개인으로부터 다양한 종류의 데이터가 생산되고 있다. 많은 사람들이 이런 데이터와 정보를 구분없이 혼용해서 사용하고 있는데, Fig. 1에서와 같이 데이터 그 자체만으로는 의미가 크지 않으며, 특정한 목적과 특징으로 체계적으로 구분하여 관리될 때 보다 큰 의미를 가질 수 있다[3,4]. 이와 같이 데이터를 효과적으로 관리하고 정보로서 이용하기 위해서는 데이터베이스(Database)라는 도구를 사용하며, 데이터베이스의 일반적인 용도는 저장이지만, 저장된 데이터의 검색과 갱신 또는 중요한 기능이라고 할 수 있다. 대량의 데이터를 더욱 쉽게 이용하기 위해서는 데이터베이스를 사용해야 한다[5].

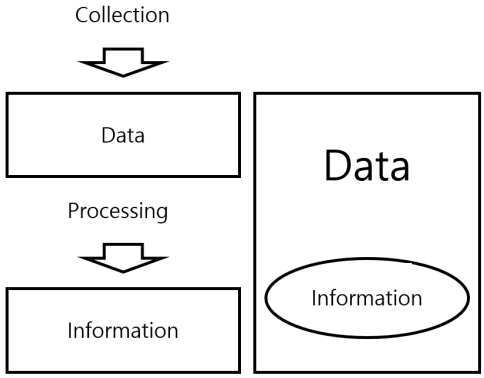

Fig. 1. Data & Information

2.2 빅데이터

최근의 4차 산업혁명시대에 빅데이터라는 단어가 주요 키워드로 등장하면서, 일반인들도 데이터보다는 빅데이터라는 용어를 자주 사용하고, 친숙하게 되었다. 빅데이터(Big data)란 글자 그대로 큰 데이터 즉, 대량의 데이터라고 할 수 있으며, 일반적으로는 Fig. 2와 같이 거대한 규모(volume), 빠른 속도(velocity), 높은 다양성(variety)을 특징으로 하는 데이터이다[6]. 이와 같이

빅데이터는 방대한 양의 데이터로 인해, 기존의 데이터베이스를 사용하여 저장 및 관리할 수 있는 범위를 넘어서므로 이를 위해 NoSQL, R 언어, Hadoop 와 같은 빅데이터 처리기술이 제시되고 있다[1,6].

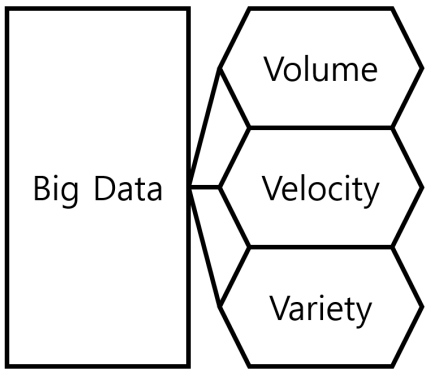

Fig. 2. Big Data Characteristics

2.3 시계열 데이터

일반적으로 데이터는 형태에 따라 정형 데이터(Structured data), 반정형 데이터(Semi-structured data), 비정형 데이터(Unstructured data)로 구분할 수 있으며, 데이터베이스에 저장되는 데이터의 양 측면에서 과거에는 정형 데이터의 비율이 높았다면, 최근에는 스마트폰의 폭발적인 보급과 더불어 소셜네트워크 서비스(SNS) 확대로 비정형 데이터의 비율이 높아지고 있다. 또한, 사물인터넷(IoT, Internet Of Things)과 헬스케어(Health care) 등의 기술개발과 활용도가 높아짐으로써 시간정보를 포함하고 있는 시계열 데이터의 중요성 또한 한층 높아졌다[7]. 이와 같이 시계열 데이터란 일정 기간에 대해 시간의 함수로 표현되는 데이터를 말하며, 과거 몇 년간의 월별 매출액, 일변 주식시세 등과 같은 예가 시계열 데이터라고 할 수 있다. Fig. 3은 2021년 1월부터 2023년 8월까지 KOSPI의 주가지수 즉, 시계열 데이터를 나타내는 그래프이다. 그래프를 살펴보면, 시작시점인 2021년 1월부터 KOSPI의 주가지수가 마지막 시점인 2023년 8월까지 월별로 어떻게 변화되고 있는지 시간적 추이를 알 수 있다.

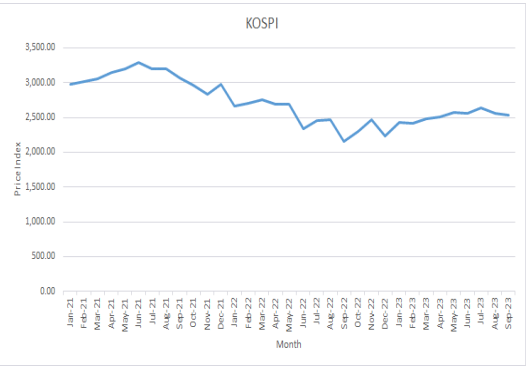

Fig. 3. Time Series Data

과거에는 이런 환경에서 수집되는 데이터를 저장한 다음에, 특정한 상황 발생시 해당 데이터를 분석하는 용도로만 이용되었지만, 최근에는 인공지능 기술이 더욱 발전함으로써 시계열 데이터를 활용한 미래 예측 분야에서 활용이 크게 증가하였다.

3. 시계열 데이터 활용 현황

3.1 시계열 데이터와 변동요인

2장에서 기술한 대로 시계열 데이터는 일정 시간 간격으로 측정된 데이터로서, 환율, 유가, 취업률, 주가지수, 연중기온 등과 같이 우리 일상생활 주변에서 자주 접하고 있는 데이터이다. 예시와 그림 3의 그래프에서도 알 수 있듯이 시계열 데이터는 시간 축을 따라 변화되는 동적 데이터라고 할 수 있다. 또한, 시계열 데이터는 오랜 기간 동안 수집되어야만 의미가 있을 뿐만 아니라 이를 활용한 결과 또한 가치가 있다[8].

시계열 데이터는 순환(cycle factor), 계절(seasonal factor), 추세(trend factor), 불규칙(irregular factor)과 같은 4가지 성분으로 구성된다[9-12]. 순환 요인은 시간의 흐름에 따라 상하로 반복되는 변동으로 추세선을 따라 변화하는 것으로, 시장 상황이나 정치, 경제와 사회적 이슈에 따른 순환적 변화를 감지한다. 계절 요인은 수집한 시계열 데이터들을 연 단위나 그보다 작은 단위인 분기나 월별 주기로 나타냈을 때 자연조건, 사회적 관습 등의 영향을 받아서 계절별로 차이를 보이는 것이다. 추세 요인은 장기적으로 증가하거나, 감소하는 경향이 보이는 것으로서, 데이터가 어떤 형태를 취하는 것이다. 마지막으로 불규칙 요인은 명확히 설명될 수 없는 요인에 의해 발생되는 것으로서, 자연재해와 같이

사전에 예상할 수 없는 특별한 사건으로 인해 발생한다. 근로현장에의 파업도 불규칙 요인의 대표적인 예라고 할 수 있다.

3.2 시계열 데이터 분석

우리 민족은 1441년에 세계에서 가장 먼저 강우량을 측정하는 측우기를 발명하여 전국적인 강우량 관측망을 구축하였으며, 전 세계적으로 과거로부터 기상정보를 수집하여 20세기에 들어서는 일기예보를 위한 컴퓨터 시스템을 구축하였으나, 기상 예측을 위한 초기 시도는 정확한 결과를 예측하지 못하였다. 이의 원인으로는 모든 자연현상을 한 번에 고려할 경우의 수가 너무 많은데 기인하였다[7].

시계열 데이터 분석은 주어진 데이터의 패턴을 통해 시간적 인과관계를 찾는 과정이라고 정의할 수 있으며[13], 시간에 종속적인 관계를 가지므로 이를 분석에 사용함으로써 시계열 데이터 분석을 보다 정확하게 할 수 있다. 또한, 시계열 데이터를 분석하는 목적은 수집한 데이터를 발생시키는 확률적 체계를 이해하고 모형화하는 것과, 수집된 과거의 데이터를 통해 미래의 값을 예측하는데 있다[14]. 이처럼 시계열 데이터의 종류에 따라 분석 방법도 다르게 적용되어야만 정확한 분석결과를 도출할 수 있다.

투입된 변수의 수에 따라 일변량과 다변량으로 나눌 수 있으며, 단일 시간 종속 변수만을 사용하는 일변량 분석방법으로는 Box-Jenkins 방법, 지수평활 방법, 시계열 분해방법이 일반적이고, 시간당 접속 사용자의 수, 도시의 날짜별 온도 등이 일변량 시계열에 해당한다. 시계열 데이터와 함께 설명변수가 있는 경우인 다변량 분석방법으로는 계량경제 모형, 전이함수모형, 개입분석, 상태공간 분석, 다변량 ARIMA 등이 있으며, 기업의 분기별 재정 안정성은 회사의 수입, 부채 등의 여러 종속 변수들을 이용하므로 다변량 시계열에 해당한다[15,16].

3.3 RNN과 LSTM

머신러닝(machine learning) 분야에서는 시계열 데이터 분석에 일반적으로 RNN(Recurrent Neural Network)과 LSTM(Long-Short Term Memory)을 사용한다. RNN은 순환신경망으로 불리며, 다층 신경망(Multilayer Neural Network)와 달리 Fig. 4와 같이 과거의 데이터를 사용할 수 있도록 내부에 순환 구조를 가진다[17]. RNN은 딥러닝 모델의 한 종류로서, 각 노드가 순환 경로를 가지므로 이전의 정보를 기억한 다음에 이것을 바로 다음 단계의 정보 즉, 입력으로 사용하는 것이 특징이며, 이런 특징을 이용하여 시컨스 데이터를 처리하는데 높은 성능을 나타낼 수 있다. RNN의 이러한 메모리 기능은 과거의 정보를 현재의 작업에 사용함으로써 미래의 정보를 예측하는데 활용될 수 있으며, 이는 과거의 정보를 기억하는 동시에 최신의 데이터 갱신이 가능함을 의미한다.

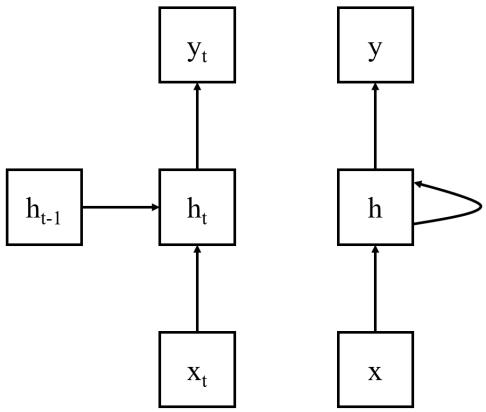

Fig. 4. The structure of RNN

Fig. 4에서 현재 상태의 Hidden h_t는 직전의 Hidden h_{t-1}을 받아서 갱신되는 형태인 수식 1과 같이 비선형함수인 하이퍼볼릭탄젠트(tanh)로 나타낼 수 있다.

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \quad (1)$$

앞서 기술하였듯이 RNN은 직전의 은닉층 상태를 다음 은닉층으로 전달함으로써 이전의 정보를 기억함으로써 시계열 데이터 분석에 효과를 높일 수 있지만, 관련 정보와 그 정보를 사용하는 시점 사이가 멀어질수록 역전파를 통해 가중치를 업데이트할 때 기울기가 점차 소실됨으로써 가중치가 정상적으로 업데이트되지 않아 학습 능력이 현저히 저하된다. 이것은 장기 문맥 의존성(long-term context dependency)이라고 부르며, 시계열 데이터의 특성상 시간상 멀리 떨어져 있다고 하더라도 이들 두 요소는 밀접한 상호작용을 함을 의미한다.

LSTM은 RNN의 기울기 소실(vanishing gradient) 문제를 보완하기 위한 모델로서, RNN에 선별 기억 능력을 추가한 신경망이라고 할 수 있다[18]. LSTM은 그림 5에서 나타내듯이 게이트를 통해 선별 기억 능력을 확보하는데, 입력, 삭제, 출력 게이트를 이용하여 정보의 흐름 조절과 더불어 필요한 정보를 기억하거나 불필요한 정보를 잊는 기능을 수행함으로써 장기 문맥의 이해할 수 있게 된다.

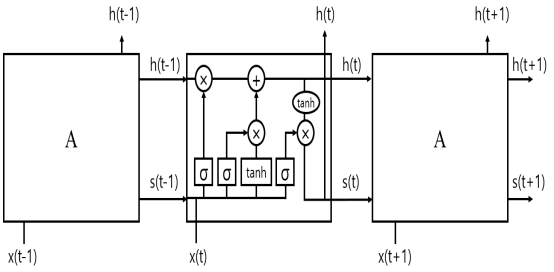

Fig. 5. LSTM module

Fig. 5에서 i번째 입력 게이트 $g_i(t)$,삭제 게이트 $f_i(t)$, 출력 게이트 $q_i(t)$는 다음과 같은 식으로 계산된다.

$$g_i(t) = \sum_j U_{ij}^g x_j(t) + \sum_j W_{ij}^g h_j(t-1) + b_i^g \quad (2)$$
$$f_i(t) = \sum_j U_{ij}^f x_j(t) + \sum_j W_{ij}^f h_j(t-1) + b_i^f$$
$$q_i(t) = \sum_j U_{ij}^q x_j(t) + \sum_j W_{ij}^q h_j(t-1) + b_i^q$$

3.4 인공지능 기법을 이용한 시계열 데이터 분석

과거에는 시계열 데이터 분석을 통해 이용하기 위해서는 ARIMA(AutoRegressive Integrated Moving Average), SARIMA(Seasonal ARIMA)를 주로 사용하였다. 최근에는 딥러닝 기술의 발전으로 RNN과 LSTM을 사용한다. 이런 모델들은 자연어 처리, 음성인식, 주식 또는 날씨 예측 분야에서 다양하게 활용되고 있다[17-19].

오종민, 신현수, 신예슬, 정형철(2017)은 서울시 미세먼지 예측을 위해 2001~2014년까지의 월별 평균 미세먼지 PM10 농도 데이터를 사용하여 여러 가지 시계열 분석법을 사용하여 비교하여 예측에 적용해보았다[18].

한편, 배성완, 유정석(2018)은 부동산 가격지수 예측을 위해 시계열 분석 모형과 다양한 머신러닝 방법을 이용한 결과 많은 조건에서 LSTM이 우수함을 확인하였다[21]. 송한진, 최흥식, 김선웅, 오수훈(2019)은 금융상품들에 대한 미래 변동성을 예측하기 위해 LSTM 기법을 활용하여 과거에 수집된 금융상품의 데이터를 학습시켜서 미래의 가격 변동성을 예측해 본 결과 같은 기간의 실제 변동성 값과 유사한 결과를 보임을 확인하였다[22].

시간 축을 따라 신호가 변하는 동적 데이터인 시계열 데이터는 다양한 시계열 분석법을 사용하여 해당 데이터에 대해 적절한 방법을 찾는 연구가 진행되었고, 인공지능 기술인 딥러닝은 시계열 데이터 처리에서 혁신을 이루었고, 의료 진단, 주식 시세 분석, 기후 예측 등 다양한 분야에 활용되어 높은 예측 결과를 보이고 있다.

4. 결론

인류의 출현과 함께 시작된 역사에는 기록이라는 수단이 있기에 현재에 사는 우리는 데이터를 통해 과거를 확인할 수 있다. 동굴 속의 벽화에서부터 최근의 4차 산업혁명에 이르기까지 데이터의 유형 또한 다양하며, 이들 데이터 중에서 시간 정보가 들어 있는 시계열 데이터도 한 종류이다. 본 논문은 수많은 데이터 중에서 시계열 데이터의 활용 동향을 알아보기 위해서 시계열 데이터의 개념에서부터 머신러닝 분야에서는 시계열 데이터 분석에 주로 사용되는 RNN과 LSTM에 대해 분석하고, 이런 모델들을 활용한 사례를 조사하였다. 경제, 기상, 금융 등 다양한 분야에서 시계열 분석과 예측을 통해 과거 동향을 이해하고 미래 값을 예측하여 의사 결정을 지원하는 데 활용되고 있다. 머신러닝 이전에는 feature 즉, 학습 및 예측을 할 데이터의 특징, 항목을 사람의 손으로 만들었지만, 딥러닝에서는 이런 작업을 모델 안에서 자동으로 해줄 뿐만 아니라, 빠른 분석을 통해 더욱 정확한 예측을 할 수 있게 되었다. 또한, 사물인터넷과 헬스케어의 기술개발과 활용도가 높아짐으로써 심전도 신호, 주식시세 분석 등 보다 다양한 분야에서 시계열 데이터의 분석이 활용될 것이다.

REFERENCES

[1] Kim. Y. H. (2022). *Introduction to Databases(3rd edition).* Seoul : Hanbit Academy.

[2] Viktor. M. S. & Kenneth. C. (2013). *The World Created by Big Data*. Seoul : 21st Century Books.

[3] Wikipedia. (n.d). https://ko.wikipedia.org/wiki/%EC%9E%90%EB%A3%8C

[4] Kim. J. Y. (2013). *Database Basics and Practice*. Seoul : Hanbit Media.

[5] Mic. & Kimura. M. (2016). *First Steps to Database*. Seoul : Hanbit Media.

[6] Oracle. (n.d). *What is Big Data?* (Online). https://www.oracle.com/kr/big-data/what-is-big-data/

[7] Aileen. N. (2021). *Practical Time Series Analysis*. Seoul : Hanbit Media.

[8] Kim. E. D., Ko. S. K., Son. S.C. & Lee. B.T. (2021). Technical Trends of Time-Series Data Imputation, *Electronics and Telecommunications Trends, 36(4).*

[9] Park. S. Y. & Moon. B. H. (2000), Similarity Search in Time-Series Databases Using Decomposition Method, *Korea Computer Congress 2000, 27(2)*, 110-112.

[10] Kim. H. W. (2004). Cyclical Analysis on the Composite Indexes of Business Indicators, *Journal of the Korean Official Statistics, 9(1)*, 29-52.

[11] Jeon. I. J. (2022). One month study-data analysis. Seoul : Discovery Media.

[12] Lee. H. Y. & Lee. P. Y. (2003). Learning statistics through stories. Paju : Jayu Academy.

[13] Jin. Y. H., Ji. S. H. & HAN. K. H. (2021). Time Series Data Analysis and Prediction System using PCA, *Journal of The Korea Convergence Society, 12(11)*, 99-107.

[14] Choi. B. S. (2001). *Univariate time series analysis*. Seoul : Segyeongsa.

[15] Kang. C. G. (2006). Comparative analysis of time series forecasting techniques, *Quarterly National Accounts, 3(26),* 80-105.

[16] Cho. S. S. & Son. Y. S. (2002). *Time series analysis using SAS/ETS*. Seoul : Yulgok Publishing.

[17] Park. H. J., Seok. K. H., Shim. J. Y. & Hwang. C. H. (2019). *Learning deep learning with TensorFlow*. Seoul : Hanbit Academy.

[18] Oh. O. S. (2021). *Artificial intelligence made with Python*. Seoul : Hanbit Academy.

[19] Shin. H. K. (2019). Time Series Forecasting on Car Accidents in Korea Using Auto-Regressive Integrated Moving Average Model-. *Journal of Convergence for Information Technology. 9(12)*, 54-61.

[20] Oh. J. M., Shin. H. S., Shin Y. S., Jeong H. C.(2017). Forecasting the Particulate Matter in Seoul using a Univariate Time Series Approach, *Journal of The Korean Data Analysis Society, 19(5)*, 2457-2468.
DOI : 10.37727/jkdas.2017.19.5.2457

[21] Bae. S. W., Yu. J. S. (2018), Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model, *Housing Studies Review, 26(1)*, 107-133.
DOI : 10.24957/hsr.2018.26.1.107

[22] Song. H. J., Choi. H. S., Kim S. W., Oh. S. H. (2019), A Study on Financial Time Series Data Volatility Prediction Method Using AI's LSTM Method, *Journal of Knowledge Information Technology and Systems 14(6)*, 665-673.
DOI : 10.34163/jkits.2019.14.6.009

최 신 형(Shin-Hyeong Choi) **[종신회원]**

- 2002년 8월 : 경남대학교 컴퓨터공학과(공학박사)
- 2003년 9월 ~ 현재 : 강원대학교 전기제어계측공학부 교수
- 관심분야 : 임베디드시스템, 사물인터넷, 정보보안
- E-Mail : cshinh@kangwon.ac.kr