Journal of The Korea Convergence Society
Vol. 12. No. 10, pp. 9-17, 2021

I-SSN 2233-4890 / e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2021.12.10.009

적대적 생성망을 이용한 부동산 시계열 데이터 생성 방안

유재필[1], 한창훈[2], 신현준[3*]
[1]키스채권평가 차장, [2]멀티에셋자산운용 상무, [3]상명대학교 공과대학 경영공학 교수

A Methodology for Realty Time-series Generation Using Generative Adversarial Network

Jae-Pil Ryu[1], Chang-Hoon Hahn[2], Hyun-Joon Shin[3*]
[1]Deputy Department Head, KIS Pricing
[2]Managing Director, Multiasset
[3]Professor, Dept. of Management Engineering, Sangmyung University

요 약 최근 빅데이터 분석, 인공지능, 기계학습 등의 발전으로 인해서 데이터를 과학적으로 분석하는 기술이 발전하고 있으며 이는 의사결정 문제를 최적으로 해결해주고 있다. 그러나 특정 분야의 경우에는 데이터의 양이 부족해서 과학적 방식에 적용하는 것이 어렵다. 예컨대 부동산과 같은 데이터는 데이터 발표 시점이 최근이거나 비 유동성 자산이다 보니 발표 주기가 긴 경우가 많다. 따라서 본 연구에서는 이런 문제점을 극복하기 위해서 TimeGAN 모형을 통해 기존의 시계열의 확장 가능성에 대해서 연구하고자 한다. 이를 위해 부동산과 관련된 총 45개의 시계열을 데이터 셋에 맞게 2012년부터 2021년까지 주 단위로 데이터를 수집하고 시계열 간의 상관관계를 고려해서 총 15개의 최종 시계열을 선정한다. 15개의 시계열에 대해서 TimeGAN 모형을 통해 데이터 확장을 한 결과, PCA 및 T-SNE 시각화 알고리즘을 통해 실제 데이터와 확장 데이터 간의 통계적 분포가 유사하다는 것을 확인할 수 있었다. 따라서 본 논문을 통해서 데이터의 과적합 또는 과소적합이라는 한계점을 극복할 수 있는 다양한 실험이 연구되기를 기대한다.

주제어 : 생성적 대립 신경망, 시계열 생성적 대립 신경망, 시계열 자료, 기계 학습, 딥 러닝, 데이터 확장

Abstract With the advancement of big data analysis, artificial intelligence, machine learning, etc., data analytics technology has developed to help with optimal decision-making. However, in certain areas, the lack of data restricts the use of these techniques. For example, real estate related data often have a long release cycle because of its recent release or being a non-liquid asset. In order to overcome these limitations, we studied the scalability of the existing time series through the TimeGAN model. A total of 45 time series related to weekly real estate data were collected within the period of 2012 to 2021, and a total of 15 final time series were selected by considering the correlation between the time series. As a result of data expansion through the TimeGAN model for the 15 time series, it was found that the statistical distribution between the real data and the extended data was similar through the PCA and t-SNE visualization algorithms.

Key Words : GAN, TimeGNA, Time-series, Machine learning, Deep learning, Data extension

*Corresponding Author : Hyun-Joon Shin(hjshin@smu.ac.kr)
Received July 7, 2021 Revised July 15, 2021
Accepted October 20, 2021 Published October 28, 2021

1. 서론

컴퓨터 기술이 지속적으로 발전하면서 정형 데이터(structured data)를 다양한 과학적 기법에 적용함으로써 다양한 의사결정 문제를 효과적으로 해결하고 있다. 대표적인 사례로 알파고(AlphaGo)가 있는데 이는 방대한 데이터를 머신러닝(machine learning)으로 스스로 학습하여 최적의 의사결정을 할 수 있게 모델링(modeling)이 되어 있다[1]. 또한 스마트폰 보급이 증가하면서 인터넷상에 축적되는 데이터의 양은 크게 증가하고 있고 소셜네트워크서비스(social network service, 이하 SNS)가 대중화 되면서 많은 비정형 데이터(unstructured data)들이 실시간으로 쌓여가고 있다[2]. 이러한 비정형 데이터가 지속적으로 저장되고 있는 SNS의 정보들을 학습 기법을 이용해서 미래를 예측하는 연구 사례가 증가하고 있다[3]. 즉 과거에 주로 회귀분석을 통해서 미래 데이터를 예측하는 것이 4차 산업 시대에 도래하면서 방대한 빅데이터와 딥 러닝(deep learning)과 같은 기법 등을 통해서 예측이 이뤄지고 있다. 그러나 빅데이터에는 그만큼 노이즈(noise) 정보도 매우 많기 때문에 우리가 원하는 의사결정 문제를 해결하기 위해서는 실효성이 있는 정보를 추출하는 것이 무엇보다 중요하다. 이처럼 데이터의 충분성과 과학 기술의 발전으로 인해서 다양한 난제를 풀 수 있는 반면에 데이터의 비충분성으로 인해서 기계학습(machine learning)과 같은 기법에 적용하기 어려운 경우도 있다. 특히 데이터의 수리적 분류(Classification)의 성과는 데이터의 양과 질에 비례할 수밖에 없으며 머신러닝과 같은 데이터 학습에 있어서도 매우 적은 정보는 물론 너무 복잡한 분포를 갖는 데이터로 학습을 시도하면 과적합(overfitting) 또는 과소적합(underfitting) 문제가 발생할 수 있다.

Fig. 1은 데이터의 특성에 따른 학습도를 도식화한 것인데 과소적합인 경우에는 데이터의 차원이 너무 단순해서 명확한 패턴(pattern)을 포착할 수 없고 과적합은 불필요한 잡음을 과도하게 모델링에 반영해서 새로운 데이터에 대한 예측력이 떨어지게 된다.

학습 기법에서 학습 데이터와 입력 데이터가 달라져도 예측력의 성과가 크게 달라지지 않는 것을 일반화(generalization)라고 하는데 이와 과적합 문제가 발생하면 일반화와 멀어지는 결과를 초래한다[4]. 과적합은 목표한 수준 이상으로 학습을 하면서 입력 데이터가 학습 모형에 들어왔을 때 예측을 못하는 것을 의미한다. 학습 데이터는 실제 데이터의 부분집합인 경우가 많아서 학습 데이터의 경우에는 오차가 감소하지만 실제 데이터에서는 오차가 증가하게 된다. 예컨대 과적합이 발생하면 밝은 색의 강아지의 특성을 학습한 모형은 검은색의 강아지를 구분하지 못한다.

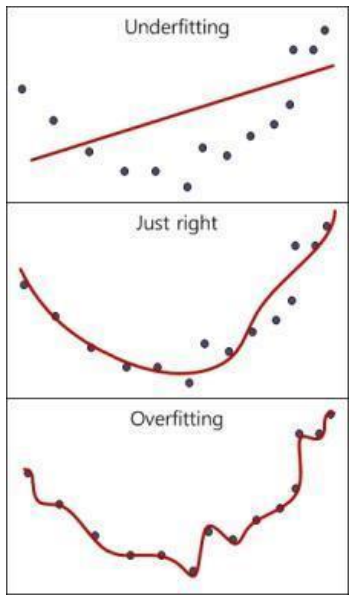

Fig. 1. Training data points by type

이런 문제를 해결하기 위해서는 입력 변수의 수를 줄이거나 학습 데이터를 늘리는 방법이 있는데 전자의 경우는 선택적이지만 후자의 경우는 실제 데이터 수가 적은 경우가 많아 이를 해결하는데 어려움이 있을 수 있다. 예로 부동산 데이터를 학습 기법을 이용해 부동산 가격을 예측하는 연구가 많이 진행되고 있는데 대부분의 연구에서 데이터 확보에 대한 한계점을 내포하고 있음을 알 수 있다[5]. 이는 데이터 증강(data augmentation)을 통해 데이터의 양을 늘리는 작업을 할 수 있으나 일반적으로 시계열의 회전(rotation)과 같은 단순한 방식을 사용한다. 그러나 이러한 방식을 통해 생성된 시계열은 실제 시계열의 속성과 차이가 있을 수 있기 때문에 이를 학습 데이터로 사용하면 예측 성능이 떨어질 수 있다. 즉 고전적인 시계열 확정 방식은 학습 기법에서 과적합 문제를 해결할 수 있으나 학습을 통한 예측 성능을 높이는 것은 한계가 있다. 따라서 본 연구에서는 전통적인 시계열 확장의 한계점을 극복하기 위해 생성적 적대 신경망(Generative Adversarial Network, 이하 GAN) 알고

리즘을 시계열 특성에 맞게 변형한 시계열 생성적 적대 신경망(Time-series Generative Adversarial Network, 이하 TimeGAN)을 통해 실제 시계열의 속성을 내포하는 확장 시계열을 만드는 방안을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 시계열 학장 및 GAN인공신경망 학습 모형에 대한 이론적 배경 및 구조를 설명하고, 3장에서는 학습을 위한 입력 변수에 대해서 그리고 4장에서는 실험 계획 및 실험 결과에 대해서 기술한다. 끝으로 5장에서는 결론을 제시한다.

2. 관련 연구

전통적인 데이터 확장은 재조정(re-scaling), 변환(translating) 그리고 왜곡(distorting) 방법을 활용한다. 여기서 데이터란 시계열 데이터는 물론 음성 데이터와 이미지 데이터 등 어떠한 형태를 수치적으로 환산할 수 있는 모든 것을 의미한다. 주로 학습 데이터가 아닌 특정 모형의 성능을 분석하기 위한 데이터 확장은 반전(reflection)을 통해 생성하는데 유재필은 금융 상품 매매 전략 모형의 성능을 분석하기 위해서 원본 데이터를 역시계열화해서 추가적인 실험 데이터를 생성하였다[6]. 학습 분야에 있어서는 실험 데이터의 부족으로 발생하는 과적합 문제를 해결하기 위해서 리샘플링(re-sampling), 부트스트랩(bootstrap) 그리고 잭나이프(jackknife)와 같은 방식으로 원본 데이터와 유사성을 갖는 확장 데이터를 생성한다[7]. 이처럼 원본 데이터를 단순하게 반복 추출하는 방식은 시계열의 복제에 가깝고 확장이라고 보는 것은 한계가 있다. 때문에 비모수 통계학 분야에서 이와 관련된 선행 연구들이 존재하는데 Cowling는 커널 밀도 추정(kernel density estimation)을 위해 밀도 함수의 대칭성에 의존해 추가적인 데이터를 생성하는 방안을 연구했다[8]. Breiman은 한정된 시계열에 특정 상수를 곱해 일차결합을 하고 회귀분석으로 새로운 데이터의 신뢰성을 높이는 방안을 제안했다[9]. Purwar는 K-평균 군집화(K-means clustering)를 이용해서 expectation step과 maximization step을 수렴하는 과정을 통해 한정된 데이터를 확장하였는데 이는 k 값과 centroid를 임의로 정해야하고 이상치(outlier)에 매우 민감하다는 단점이 있다[10]. Wu는 날씨 데이터와 금융 데이터 등에서 시계열 중간에서 발생하는 오류 값들을 재생성하기 위해서 서포트 백터 머신(support vector machine)을 이용했으나 공백이 긴 오류 값들을 생성하는 것에는 한계점을 갖고 있다.

최근에는 GAN을 통해서 과거의 많이 사용되었던 데이터 확장의 문제점을 개선하고자 하는 연구가 활발하게 진행되고 있다. GAN은 기본적으로 실제 데이터의 확률 분포에 의존해 새로운 데이터를 만들기 때문에 더 안정적인 데이터를 확보할 수 있다. 특히 DcGAN이 나오면서 이미지 복원 및 생성에 관한 기술이 발전하고 있는데 이는 손실도가 높은 이미지 데이터에서 민감한 반응을 보인다는 단점이 있는 반면에 GAN 모형을 변형한 PG-GAN은 저해상도 이미지 데이터에서 새로운 레이어(layer)를 삽입하여 최종적으로 고해상도의 이미지 데이터를 추출할 수 있다는 장점이 있다[11]. Song는 PG-GAN을 이용해서 실제 얼굴 이미지와 거의 유사한 가상의 얼굴 이미지를 생성하고 점진적 학습(progressive training)의 실효성을 입증했다[12]. 이러한 이미지 데이터의 확장과 함께 DcGAN을 변형한 WaveGAN은 소리 데이터를 학습해서 새로운 데이터를 생성하는 알고리즘이다[13]. Odena는 판별기(discriminator)의 학습 능력을 극대화하기 위해서 특징 맵(feature map)을 임의적으로 순환하여 체커보드 패턴(checkerboard artifacts) 형태로 정의되지 않도록 페이즈 셔플(phase shuffle)을 제안하였다[14]. 이처럼 GAN에 파생된 다양한 알고리즘들은 연속성상의 시계열 데이터보다는 이미지 및 소리 데이터와 같이 스팟(spot) 데이터에 특화되어 있다. 따라서 본 연구에서는 GAN을 변형한 TimeGAN 알고리즘을 통해 시계열 데이터를 확장하는 방안을 제안하고자 한다.

3. TimeGan

본 장에서는 GAN의 이론적 배경과 함께 시계열 데이터 생성을 위해 제안하는 TimeGAN 알고리즘에 대해서 설명하고자 한다.

3.1 GAN(Generative Adversarial Network)

GAN은 비지도 학습(unsupervised learning) 기법으로 생성자(generator)와 판별자(discriminator)를 교차적 학습을 통해 실제 데이터와 매우 유사한 가상의 데이터를 생성한다[15]. Fig. 1은 GAN의 기분 구조인데 생성자 G는 무작위한 잡음 z를 입력으로 하여 실제 데

이터와 비슷한 가장의 데이터 $G(z)$를 생성하도록 학습한다. 그리고 판별자 D는 x와 $G(z)$를 입력으로 하여 실제 데이터와 가상의 데이터를 판별하도록 학습한다. 즉 생성자는 실제 데이터의 분포를 학습하여 가상의 데이터가 원본 데이터와 유사할 확률을 높이고 판별자가 이들의 진위 여부를 식별할 수 있는 확률을 높이는 과정을 반복한다. 따라서 판별자와 생성자는 적대적 학습을 위해 식(1)과 같이 목적함수가 제시된다.

$$Min_G Max_D V(D,G) \\ = E_{x\sim p_{data(x)}}[\log D(x)] \\ + E_{Z\sim pz_{(z)}}[\log(1-D(G(z)))] \quad (1)$$

where, V(D,G) is value function, E is entropy, G is generator, D is discriminator, x and z is discrete random distribution

식(1)에서 $x\sim p_{data(x)}$는 실제 데이터의 확률 분포에서 특정 데이터를 샘플링(sampling)한 것이고, $Z\sim P_{z_{(z)}}$는 가우시안 분포(gaussian distribution)를 통해 노이즈에서 추출한 데이터이다. z는 잠재적 벡터(latent vector)를 의미하는데 잠재적 공간에서의 벡터를 뜻한다. 판별자인 $D(x)$는 데이터의 진위여부를 0과 1사이 값으로 나타내며 참이면 1, 거짓이면 0으로 나오며, $D(G(z))$는 G가 생성한 가상의 데이터 $G(z)$가 참이면 1, 거짓이면 0으로 출력한다. GAN은 판별 모델인 D가 목적함수 $V(D,G)$를 최대화하는 것을 목표로 하며 이를 위해 $D(x)$가 1이 되도록 하면서 실제 데이터를 참값으로 분류하도록 D를 학습한다. 또한 $(1-D(G(z)))$의 값이 1이 되기위해 $D(G(z))$은 0이 되어야 하는데 이는 데이터의 거짓 여부를 정확하게 학습하기 위함이다. 생성모델인 G는 최소화가 될 수 있도록 학습되어야 하는데 이를 위해서는 $(1-D(G(z)))$가 0의 값이 나와야 하며 $D(Gz))$는 1이 나와야 한다. 즉 D가 참값으로 산출되도록 하는 가상의 데이터를 생성할 수 있게 G를 적대적 경쟁자로서 학습시키는 과정의 반복이 GAN의 기본적 이론이다.

Fig. 2는 앞서 설명한 GAN의 구조를 보여주고 있는 그림인데 생성자는 최대한 실제 데이터와 비슷한 가상의 데이터를 생성하는 네트워크 구조를 갖추고 있으며, 판별자는 가상의 데이터가 들어왔을 때 이것이 가상의 데이터라고 분류할 수 있는 능력을 지니게 함으로써 생성자는 더욱 실제 데이터와 유사한 가상의 데이터를 생성하게 되고 판별자는 더욱 정확하게 실제 데이터와 가상의 데이터를 분류하는 과정을 학습하게 된다.

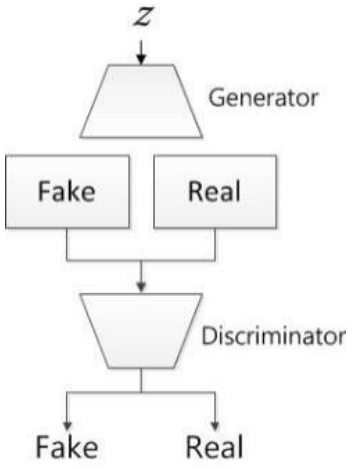

Fig. 2. The basic structure of GAN

3.2 TimeGAN

TimeGAN은 시계열 데이터가 가진 시간적 상관관계를 충분하게 반영하지 못하는 점을 개선하고 신경망의 역학을 더욱더 세밀하게 제어함으로써 현실적인 시계열 데이터를 생성하기 위한 프레임워크(framework) 설계가 가능하다[16]. TimeGAN은 총 4가지 신경망 구성 요소인 임베딩 함수(embedding function), 복구 함수(recovery function), 시퀀스 생성기(sequence generator) 그리고 시퀀스 판별기(sequence discriminator)로 구성된다. 임베딩 신경망은 잠재적 공간을 만들고 적대적 신경망은 해당 공간 내에서 작동하며 실제 데이터와 생성된 데이터를 학습을 통해 서로 간의 일치성을 높이는 구조이다.

3.2.1 임베딩 및 복구 함수

임베딩 및 복구 함수들은 적대적 신경망을 통해 데이터의 시간적 역학을 학습할 수 있도록 한다. 식(2)에서 임베딩 함수인 $e:S\times \Pi_t X\rightarrow H_S\times \Pi_t H_X$가 잠재적 코드 $h_S,h_{1:T}=e(s,X_{1:T})$에 대한 정적 및 시간적 기능들을 갖게 되며, e의 반복 신경망을 통해서 구현한다.

$$h_S=e_S(s),\ h_t=e_X(h_S,h_{t-1},X_t) \quad (2)$$

이때 $e_S : S \rightarrow H_S$는 정적 기능의 임베딩 신경망이며 $e_X : H_S \times H_X \times X \rightarrow H_X$는 시간적 기능의 반복 임베딩 신경망이다. 여기서 식(3)과 같이 각 단계별 순전파(feedforward) 신경망을 통해서 r을 구현한다.

$$\tilde{s} = r_s(h_s), \quad \tilde{X}_t = r_x(h_t) \tag{3}$$

더불어 $r_S : H_S \rightarrow S$와 $r_X : H_X \rightarrow X$는 정적 및 시간적 임베딩의 복구 신경망을 의미한다. 임베딩 및 복구 함수는 선택한 아키텍처(architecture)에 의해 매개 변수가 될 수 있으며 인과적 순서를 준수한다.

3.2.2 시퀀스 생성기 및 판별기

생성기는 합성 출력 값을 생성하기에 앞서 임베딩 공간으로 출력을 수행한다. 임베딩 및 복구 함수와 유사한 프레임워크로 $g : Z_S \times \Pi_t Z_X \rightarrow H_S \times \Pi_t H_X$는 잠재적 코드 $\hat{h}_S, \hat{h}_{1:T} = g(z_S, z_{1:T})$에 대한 정적 및 시간적 랜덤 벡터(random vectors)를 식(4)와 정의한다.

$$\hat{h}_S = g_S(z_S), \quad \hat{h}_t = g_X(\tilde{h}_S, \hat{h}_{t-1}, z_t) \tag{4}$$

판별기 또한 임베딩 공간에서 수행되는데 판별 함수인 $d : H_S \times \Pi_t H_X \rightarrow [0,1] \times \Pi_t [0,1]$는 정적 및 시간적 코드를 수신하며 $\tilde{y}_S, \tilde{y}_{1:T} = d(\tilde{h}_S, h_{1:T})$를 식(5)와 같이 반환한다. 또한 식(5)에서 $\vec{u}_t$는 각각 전방 및 후방 은닉(hidden) 상태의 시퀀스를 의미한다.

$$\tilde{y}s = d_S(\tilde{h}_S), \quad \tilde{y}_t = d_X(\vec{u}_t, \overleftarrow{u}_t) \tag{5}$$

Fig. 3은 앞서 설명한 관련 식들의 연관 관계를 보여주고 있으며 Fig. 4는 학습 계획을 보여주고 있는데 실선은 데이터의 순방향 전파를 나타내고 점선은 기울기의 역방향 전파를 나타낸다.

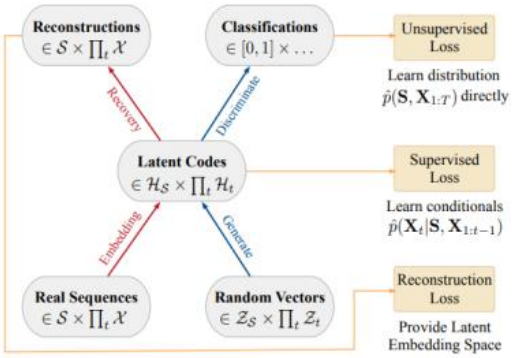

Fig. 3. Block Diagram[16]

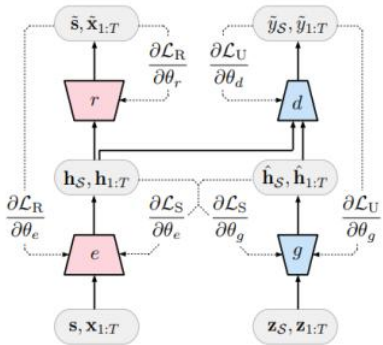

Fig. 4. Training Scheme[16]

4. 실험계획 및 분석

본 장에서는 확장할 시계열 종류 및 실험 데이터 정의 등의 실험계획을 설명하고 TimeGAN을 이용해 생성된 데이터의 통계적 일치성 여부 등을 기술한다.

4.1 실험계획

Table 1은 본 연구의 실험계획을 정리한 표이다. 총 45개의 시계열에 대해서 2012년부터 2021년도까지 주 단위로 데이터를 수집한다. 부동산과 밀접한 관련이 있는 시계열들은 일별보다는 주 별 또는 월 별 데이터들이 대부분이기 때문에 시계열 간의 데이터 셋(data set)을 통일하기 위해 주 별로 설정한다. 또한 앞서 설명했듯이 상관관계를 통해 최종 18개의 시계열의 실제 데이터와 확장 데이터간의 유의적 관계를 설명하기 위한 고차원 데이터의 시각화 알고리즘을 구연하기 위해 본 연구에서는 PCA(Principal Component Analysis)와 t-SNE(Stochastic Neighbor Embedding)을 사용한다.

Table 1. Experimental plan

Experimental Factors	Details
Time series to experiment	Table 1. Reference
collection period	2012.05~2021.02
collection cycle	Weekly
Visualization Algorithm	PCA, T-SNE
Learning Methods	TimeGAN

4.2 결과분석

시계열 확장의 대상을 선정하기 위해서 총 45개의 부동산 관련 시계열 데이터를 수집하고 실험에서의 자원(resource) 효율성을 높이고자 데이터 간의 상관관계를 산출한다. 산출된 상관계수를 바탕으로 최종적으로 시계열의 종류를 선정하였는데 이는 Table 2와 같다. Fig. 5는 45개의 시계열과 최종 선정된 18개의 시계열에 대해 시계열 간의 상관관계의 강도를 시각화한 그림인데 이는 가로 및 세로 각각 45개와 18개로 구성된 시계열별 픽셀(pixel)에 상관관계가 높아질수록 진한 색으로 표현이 되고 있다. 상단의 그림 같은 경우는 많은 부분에서 시계열 간 상관관계가 높아서 중복성이 강한 시계열이라는 것을 확연하게 알 수 있는 반면에 총 18개로 구성된 하단 그림의 경우에는 시계열 간의 통계적 성향이 상단의 그림보다는 중복적이지 않는 것을 확인할 수 있다.

Table 2. Final Time Series Data

Time Series Data Types	
Price Index	KOSPI Volume
Lease Price Index	US Bond 10y
Trade Volume	Exchage Rate
Lease Volume	WTI
Seoul Index	CRB
Seoul Build Ups	Baltic Dry
Construction Index	Bank Deposit
KOSPI	Stock Fund
Stock Customer Deposit	Goldman Raw Material

고차원적인 데이터의 경우에는 데이터의 특정 변수에 대한 분포 및 상관계수 등을 통해서 데이터의 형태를 수리적으로 분석하기 힘들다. 따라서 본 실험에서는 핵심정보를 선정하고 고차원적인 데이터를 시각화할 수 있는 PCA 및 T-SNE 알고리즘을 통해 확장 데이터의 성과를 분석하고자 한다. PCA의 경우에는 데이터의 정보를 유지하면서 데이터 차원을 축소하는 방식이며, t-SNE는 실제 데이터를 최적으로 표현해줄 수 있도록 데이터 차원을 축소하는 방식을 의미한다. 다만 t-SNE 알고리즘의 경우에는 데이터의 수의 배 이상만큼의 데이터 입력 과정이 필요하기 때문에 시각화 과정에서 PCA 보다 소요시간이 길다는 단점이 있다.

Fig. 6은 각각 PCA와 t-SNE 알고리즘을 통해서 실제 데이터와 확장 데이터 간의 통계적 분포를 시각화한 결과를 보여주고 있다. 각 3번으로 나눠 학습을 하였는데 2가지 알고리즘 모두 원본 및 확장 데이터 간에 유의한 일치성을 보이고 있다. 특히 t-SNE의 3번째 실험의 경우에는 실제 데이터와 확장 데이터의 통계적 분포가 거의 일치한 것을 확인할 수 있다. 본 그림의 경우에는 실제 데이터와 TimeGAN으로 생성된 가상 데이터간의 통계적 분포를 시각적으로 쉽게 볼 수 있게 나타낸 그림인데 실제 데이터와 가상 데이터의 타점의 일치도가 높을수록 실제 데이터를 잘 대변하고 있다는 것을 의미한다.

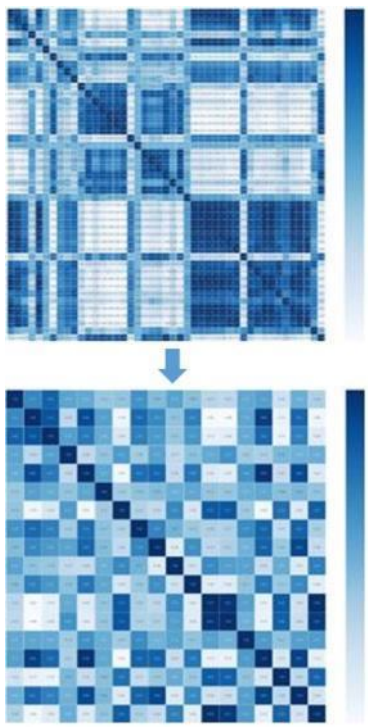

Fig. 5. Correlation Between Time Series

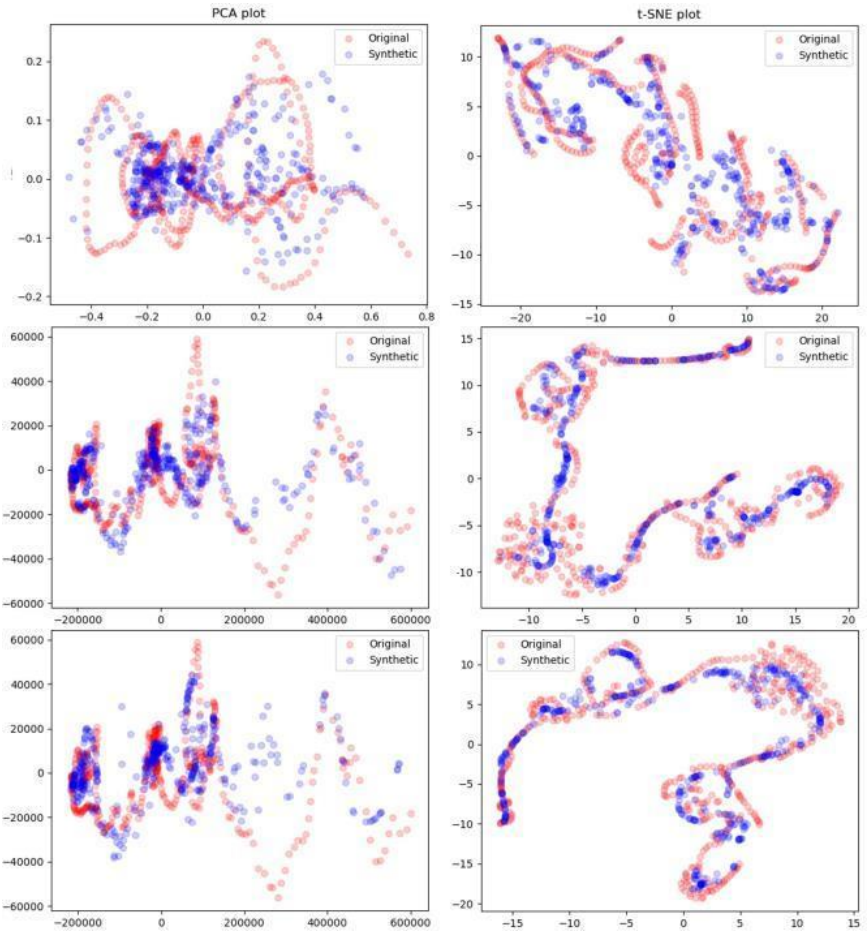

Fig. 6. Experimental results with PCA and t-SNE

5. 결론

본 논문은 특정 분야의 데이터 부족 현상으로 인해 기계학습 및 딥러닝과 같은 과학적인 의사결정 문제에 적용하는 것에서 한계가 있는 점을 극복하기 위해서 실제 데이터와 유사한 시계열을 갖는 데이터를 확장하는 가능성에 대해서 연구하였다. 이는 동시에 데이터의 과적합 및 과소적합 문제를 해결할 수 있는 대안으로서 기존의 GAN 모형을 변형한 TimeGAN 모형을 제안하였다. 전통적인 GAN 모형의 경우에는 시각적인 데이터 수치의 복재 및 확장에서는 강한 성능을 보인 반면에 연속적인 시계열 데이터의 확장에서는 많은 한계점을 내포하고 있다. 따라서 본 연구에서는 시계열 확장에 보다 더 높은 성능을 보일 수 있는 TimeGAN 모형을 제안하고 이를 통해서 원본 시계열 데이터를 확장하는 실험을 진행하였다. 실험을 위해 총 45개의 시계열 데이터를 수집하고 상관관계를 통해 최종적인 15개의 시계열 데이터를 선정하였다. 또한 실제 데이터와 학장 데이터의 유사성을 정량적으로 분석하기 위해서 고차원적 데이터의 시각적 표현 알고리즘인 PCA와 t-SNE를 통해서 통계적 유의성을 확인하였다. 그 결과 3번의 학습 모두 2개의 알고리즘에서 확실한 유사성을 보이는 것을 알 수 있었다.

본 논문은 다양한 데이터 분석 기법의 발전에 반해 데이터의 부족으로 인해 의사결정 문제를 해결하는데 어려울 수 있는 한계점을 해결하기 위한 방안으로 시계열 데이터의 확장에 대해 연구하였다. 향후 다양한 시계열 확장에 관한 연구에 있어서 본 논문이 긍정적인 참고자료가 되기를 기대한다.

REFERENCES

[1] W. K. Kang & B. R. Kim. (2019). Consideration of Human Emotions about Artificial Intelligence - Focused on the Analysis of Newspaper Articles on AlphaGo VS Lee Sedol, *Journal of Ethics, 1(132)*, 191-201.
DOI : 10.15801/je.1.132.201812.181

[2] J. P. Ryu, C. H. Han & H. J. Shin. (2016). Sector Investment strategies Using Big Data Trends, *Journal of Information Technology and Architecture, 13(1)*, 111-121.

[3] J. Y. Yim & B. Y. Hwang. (2014). Predicting Movie Success based on Machine Learning Using Twitter, *KIPS Transactions on Software and Data Engineering, 3(7)*, 263-270.

[4] S. S. Shin, H. Y. Cho & Y. H. Kim. (2021). Optimal Ratio of Data Oversampling Based on a Genetic Algorithm for Overcoming Data Imbalance, *Journal of the Korea Convergence Society, 12(1)*, 49-55.
DOI : 10.15801/je.1.132.201812.181

[5] S. W. Bae & J. S. Yu. (2018). Estimating the Real Estate Price Index Based on Sample House Price: Focusing on the Use of Machine Learning Method, *Housing Studies, 26(4)*, 53-74.
DOI : 10.24957/hsr.2018.26.4.53

[6] J. P. Ryu & H. J. Shin. (2012). Investment Strategies for KOSPI200 Index Futures Using VKOSPI and Control Chart, *Journal of the Korean Institute of Industrial Engineers, 38(4)*, 237-243.
DOI : 10.7232/JKIIE.2012.38.4.237

[7] J. W. Kim. (2019). Predictive Optimization Adjusted With Pseudo Data From A Missing Data Imputation Technique, *Journal of the Korea Academia-Industrial cooperation Society, 20(2)*, 200-209.
DOI : 10.5762/KAIS.2019.20.2.200

[8] A. Cowling & P. Hall. (1996). On pseudo data methods for removing boundary effects in kernel density estimation, *Journal of the Royal Statistical Society, 58(3)*, 551-563.
DOI : 10.1111/j.2517-6161.1996.tb02100.x

[9] L. Breiman. (1998). Using convex pseudo-data to increase prediction accuracy, *breast, 5(2)*, 1-18.

[10] A. Purwar & S. K. Singh. (2015). Hybrid prediction model with missing value imputation for medical data, *Expert Systems with Applications, 42(13)*, 5621-5631.
DOI : 10.1016/j.eswa.2015.02.050

[11] J. H. Yoon, B. K. LEE & B. W. Kim. (2021). A Study on GAN Algorithm for Restoration of Cultural Property, *Journal of The Korea Society of Computer and Information, 26(1)*, 77-84.
DOI : 10.9708/jksci.2021.26.01.077

[12] U. Sivarajah, M. M. Karnal, Z. Irani & V. Weerakkody. (2017). Critical analysis of Big Data challenges and analytical methods, *Journal of Business Research, 70*, 263-286.
DOI : 10.1016/j.jbusres.2016.08.001

[13] M. K. Back, S. W. Yoon, S. B. Lee & K. C. Lee. (2020). Improving Fidelity of Synthesized Voices Generated by Using GANs, *KIPS Trans. Softw. and Data Eng, 10(1)*, 9-18.
DOI : 10.3745/KTSDE.2021.10.1.9

[14] A. Odena, V. Dumoulin & C. Olah. (2016). Deconvolution and checkerboard artifacts, *Distill, 1(10)*, 1-3.
DOI : 10.23915/distill.00003

[15] I. Goodfellow, P. A. Jean, M. Mirza, B. Xu, W. F. David, S. Ozair, A. Courville & Y. Bengio. (2014). Generative adversarial nets, *In Advances in neural information processing systems*, 2672-2680.
DOI : https://dl.acm.org/doi/10.5555/2969033.2969125

[16] J. S. Yoon & D. E. Jarrett. (2019). Time-series Generative Adversarial Networks, *33rd Conference on Neural Information Processing System*.

유 재 필(Jae Pil Ryu) [정회원]

· 2017년 2월 : 상명대학교 일반대학원 공과대학 경영공학과(공학박사)
· 2016년 11월 ~ 현재 : KIS채권평가 평가본부 주식파생실 차장
· 관심분야 : 금융공학, 기계학습, 딥러닝
· E-Mail : jaepilryu@kispricing.com

한 창 훈(Chang-Hoon Hahn) [정회원]

· 2018년 2월 : 상명대학교 일반대학원 공과대학 경영공학과(공학박사)
· 관심분야 : 금융공학, 데이터마이닝
· E-Mail : hahn65@gmail.com

신 현 준(Hyun-Joon Shin) [정회원]

· 2002년 2월 : 고려대학교 산업공학과 (공학박사)
· 2002년 5월 ~ 2004년 4월 : 미국 Texas A&M대학교 연구원
· 2004년 6월 ~ 2005년 2월 : (주)삼성 전자 책임연구원
· 2005년 3월 ~ 현재 : 상명대학교 경영공학과 교수
· 관심분야 : 금융공학, 조합최적화 응용, 데이터마이닝
· E-Mail : hjshin@smu.ac.kr