ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 September: 25(3): 43~62 http://dx.doi.org/10.13088/jiis.2019.25.3.043

적대적 생성 모델을 활용한 사용자 행위 이상 탐지 방법

최남웅
연세대학교 산업공학과
(*nwc1114@yonsei.ac.kr*)

김우주
연세대학교 산업공학과
(*wkim@yonsei.ac.kr*)

한때, 이상 탐지 분야는 특정 데이터로부터 도출한 기초 통계량을 기반으로 이상 유무를 판단하는 방법이 지배적이었다. 이와 같은 방법론이 가능했던 이유는 과거엔 데이터의 차원이 단순하여 고전적 통계 방법이 효과적으로 작용할 수 있었기 때문이다. 하지만 빅데이터 시대에 접어들며 데이터의 속성이 복잡하게 변화함에 따라 더는 기존의 방식으로 산업 전반에 발생하는 데이터를 정확하게 분석, 예측하기 어렵게 되었다. 따라서 기계학습 방법을 접목한 SVM, Decision Tree와 같은 모형을 활용하게 되었다.

하지만 지도 학습 기반의 모형은 훈련 데이터의 이상과 정상의 클래스 수가 비슷할 때만 테스트 과정에서 정확한 예측을 할 수 있다는 특수성이 있고 산업에서 생성되는 데이터는 대부분 정답 클래스가 불균형하기에 지도 학습 모형을 적용할 경우, 항상 예측되는 결과의 타당성이 부족하다는 문제점이 있다. 이러한 단점을 극복하고자 현재는 클래스 분포에 영향을 받지 않는 비지도 학습 기반의 모델을 바탕으로 이상 탐지 모형을 구성하여 실제 산업에 적용하기 위해 시행착오를 거치고 있다.

본 연구는 이러한 추세에 발맞춰 적대적 생성 신경망을 활용하여 이상 탐지하는 방법을 제안하고자 한다. 시퀀스 데이터를 학습시키기 위해 적대적 생성 신경망의 구조를 LSTM으로 구성하고 생성자의 LSTM은 2개의 층으로 각각 32차원과 64차원의 은닉유닛으로 구성, 판별자의 LSTM은 64차원의 은닉유닛으로 구성된 1개의 층을 사용하였다.

기존 시퀀스 데이터의 이상 탐지 논문에서는 이상 점수를 도출하는 과정에서 판별자가 실제데이터일 확률의 엔트로피 값을 사용하지만 본 논문에서는 자질 매칭 기법을 활용한 함수로 변경하여 이상 점수를 도출하였다. 또한, 잠재 변수를 최적화하는 과정을 LSTM으로 구성하여 모델 성능을 향상시킬 수 있었다. 변형된 형태의 적대적 생성 모델은 오토인코더의 비해 모든 실험의 경우에서 정밀도가 우세하였고 정확도 측면에서는 대략 7% 정도 높음을 확인할 수 있었다.

주제어 : 오토 인코더, 이상 점수, 잠재 변수 최적화, 자질 매칭, 이상 탐지 적대적 생성 신경망

논문접수일 : 2019년 6월 4일 논문수정일 : 2019년 7월 22일 게재확정일 : 2019년 8월 2일
원고유형 : 학술대회(급행) 교신저자 : 김우주

1. 서론

이상 탐지는 통계적 방법과 기계 학습 방법으로 분류할 수 있다. 과거의 이상 탐지 방법은 통계적 기법을 통해 데이터 분포의 분산과 평균을 추정해 이상을 검출하는 Shewhart chart, CUSUM (Cumulative Sum), EWMV(Exponentially Weighted Moving Variance)등 (Sun et al., 2014; Vilbert et al., 2003; Nong et al., 2001)의 모델을 사용하였다. 하지만 빅데이터 시대의 도래로 데이터의 속성이 복잡해지고 양이 방대해 짐에 따라서 성능적인 측면에서 고전적 방법은 한계에 봉착하게

되었다. 이러한 점을 극복하고자 연구자들은 기계 학습을 이상 탐지에 접목하려는 연구를 진행하여 지도 학습 기반의 이상 탐지론을 사용하게 되었다.

지도 학습 방법은 정답이 주어져 있는 데이터를 바탕으로 학습한 뒤, 이상을 탐지하는 모델을 말하며 보통 SVM, 로지스틱 회귀 등을 활용한다(Shon et al., 2005). 하지만 “이상”이라고 정의되는 데이터는 실제 산업에서 잘 발생하지 않는 사건이기 때문에 이상 탐지 분야의 분석 대상 데이터는 항상 클래스 불균형 문제를 동반하고 있다. 클래스가 불균형한 데이터를 지도 학습 기반의 분류 모델에 학습시킨다면 일률적인 예측 값을 도출하게 되며 이는 정답의 타당성을 결여하는 요인이 된다. 따라서 지도 학습 기반의 이상 탐지 모델은 현실에서 적용하기에 적절하지 않다.

한편, 비지도 학습 기반의 이상 탐지 방법은 정답이 주어지지 않는 데이터를 기반으로 해당 데이터가 이상인지 정상인지 예측할 수 있는 방법을 제시하며 정답 클래스의 분포에 관계없이 합리적인 예측을 수행한다. 근래에 비지도 학습이 각광받는 이유는 앞서 언급한 지도 학습 모델의 한계점을 생각해 볼 때, 자연스러운 수순이라고 말할 수 있다. 본 연구의 핵심 모델도 이러한 추세에 발맞춰 비지도 학습 기반의 모델인 적대적 생성 신경망(Generative Adversarial Networks)을 사용하여 연구를 진행하였다.

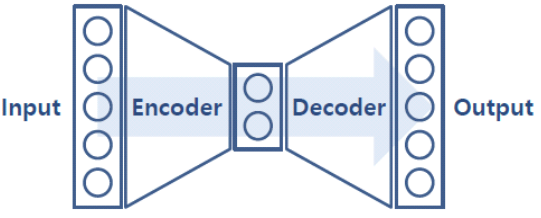
적대적 생성 신경망은 이미지 데이터 또는 영상을 다루는 도메인에서 많이 연구되고 있지만(Thomas et al., 2017; Mahdyar et al., 2017), 시퀀스 데이터를 활용하는 연구는 미비한 실정이다. 그러므로 이미지나 영상 분야에 특화된 적대적 생성 신경망을 본 연구의 분석 도메인인 시퀀스 데이터에 사용하기 위해 구조적인 변경을 모색하고, 비슷한 종류의 비지도 학습 모델 오토인코더와 성능을 비교, 분석하여 적대적 생성 모델의 강점과 특징을 본 논문을 통해 조명하고자 한다.

논문 전체의 구성은 2장에서 분석 도메인 및 기존의 개발된 이상 탐지의 선행 연구를 설명하고 3장에서는 데이터의 처리방법, 4장에서는 연구에 사용된 모델의 설명과 기존 연구와의 차별성을 서술할 것이다. 5장에서는 성능 결과를 확인하여 6장 결론부에서 본 실험의 결과로부터 적대적 생성 신경망의 강점과 활용 범위 및 본 논문의 차별성으로 인한 결과의 변화를 고찰한다. 최종적으로 7장에서는 향후 연구의 방향성 및 한계점을 언급하며 논문을 마무리할 것이다.

2. 관련 연구

2.1 오토인코더

오토인코더(Autoencoder)는 수학에서의 항등함수와 비슷한 특성을 가진 비지도 학습 기반의 신경망 모델을 말한다. 즉, 모델의 출력 값을 입력 값의 근사치로 만들며 전체 구조는 <Figure 1>과 같다.



〈Figure 1〉 Autoenocder

조금씩 차이가 있지만, 학습과정에서 목적함수는 보통 MSE(Mean Squared Error)를 사용한다. MSE는 거리 개념의 척도로써 두 개체간의 차이를 정량적으로 확인할 수 있는 값을 의미하고 이를 수식으로 나타내면 아래의 (1)과 같다. x_i는 실제 데이터, $f(\cdot)$는 인코더, $g(\cdot)$는 디코더, $g(f(x_i))$는 모델로부터 생성된 데이터를 의미한다.

$$L\left(x_i, g\left(f(x_i)\right)\right) = ||x_i - g\left(f(x_i)\right)||^2 \quad (1)$$

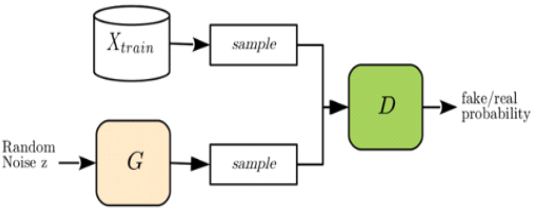
$$\hat{\theta} = argmin_\theta \sum_{i=1}^{n} L\left(x_i, g\left(f(x_i)\right)\right) \quad (2)$$

(2)와 같이 MSE을 최소화하는 θ(오토인코더의 모수)를 학습한다면 오토인코더는 실제 데이터를 모사하는 출력 값을 생성하게 된다. 역전파(Backpropa)를 통해 업데이트된 인코더는 현 데이터의 차원에서 낮은 차원으로 입력 값을 보낼 때, 핵심 자질(Feature)만을 추출하게 되고 결과적으로 주성분 분석과 유사하게 활용될 수 있다(Hinton et al., 2006). 또한, 길이나 크기가 다른 시퀀스 데이터를 제로 패딩(Zero Padding)하여 사용자 지정 크기에 은닉층의 잠재 벡터(Latent Vector)로 재 표현(Representation)함으로써 입력 값의 크기를 통일할 때, 사용되기도 한다.

이상 탐지 분야에서의 오토인코더는 “정상”이라고 정의된 데이터만을 사용하여 학습한 뒤, 이상 탐지 시점에서 입력과 출력의 차이로 발생되는 재건 손실 값(Reconstruction Loss)을 통해 이상 유무를 판단한다(An et al., 2015). 즉, 비정상의 데이터가 입력되면 정상의 데이터만을 생성하도록 학습된 디코더에 의해서 큰 손실 값이 발생하게 되고 특정 크기 이상의 손실 값 분기점(threshold)을 기준으로 데이터의 정상, 비정상을 구분한다는 것이다. 여기서 재건 손실 값이란 맨하탄 거리($|A-B|$), 유클리디안 거리($\sqrt{(A-B)^2}$) 등의 거리 척도를 사용하여 도출되는 값으로 본 논문에서는 유클리디안 거리와 비례관계에 있는 MSE값 자체를 재건 손실 값으로 적용하였다.

2.2 적대적 생성 신경망

적대적 생성 신경망은 생성자(Generator)와 판별자(Discriminator)로 구성된 네트워크를 말하며(Goodfellow et al., 2014) <figure 2>는 적대적 생성 신경망의 구조를 보여준다.



〈Figure 2〉 Generative Adversarial Networks

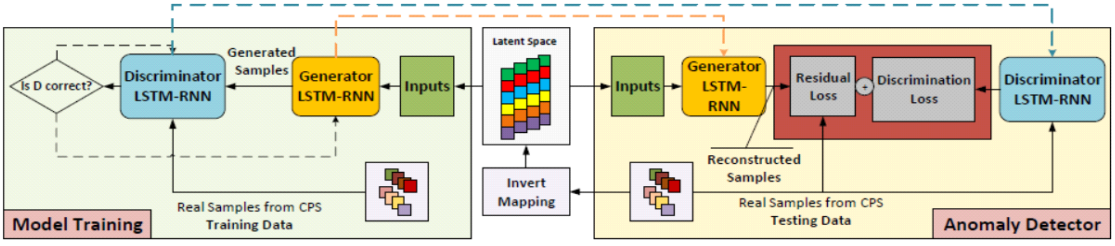
생성자는 잠재 공간(Latent space)에 실제 데이터의 분포를 매핑해서 그 분포로부터 도출된 변수를 받아 정교한 위조데이터를 생성하고 판별자는 생성된 위조데이터와 실제데이터를 구별하는 역할을 한다. 적대적 생성 신경망은 게임이론 형태의 목적 함수를 사용하여 두명의 플레이어(생성자와 판별자)가 서로 경쟁하면서 균형점(Nash equilibrium)을 찾아가는 방식으로 학습이 이루어진다. 아래 (3)은 적대적 생성 신경망의 목적함수를 나타내고 $D(x)$는 판별자가 도출하는 실제 데이터일 확률, $G(z)$는 생성자가 잠재 변수를 받아 생성한 위조 데이터, $D\left(G(z)\right)$는 생성된 위조데이터의 실제데이터일 확률을 의미한다.

$$min_G max_D V(D,G) = E_{x\sim P_{data}(X)}[logD(x)] + E_{Z\sim P_z(Z)}[\log\{1 - D(G(z))\}] \quad (3)$$

기본적인 적대적 생성 신경망은 학습하고 위조데이터를 생성하는 과정에서 다양한 문제점이 존재한다. 첫번째로 목적 함수 형태를 주목해보자. 컴퓨터는 minmax와 maxmin의 순서를 고려하지 않기 때문에 생성자의 입장에서 실제데이터의 다양성을 반영하지 않고 전체 목적 함수의 값을 낮추는 단일한 위조데이터를 생성할 수 있다. 즉, maxmin의 경우로 적대적 생성 신경망이 최적화 되었을 때, 여러 모드가 존재하는 데이터 분포를 단일 모드의 분포 형태로 학습이 되는 Mode Collapse 문제가 발생할 수 있다는 것이다. 두번째는 서로 경쟁하면서 학습되는 목적 함수의 특성상 생성자와 판별자의 학습률이 조금만 달라져도 힘의 균형이 무너져 한쪽이 학습되지 않는 문제가 존재한다. 세번째는 잠재 변수(Latent Variable)가 연속형 가우시안 분포를 따른다고 가정하기 때문에 범주형 도메인에서 실제로 존재하지 않는 범주 사이의 데이터가 생성될 수 있다는 점이다. 본 연구에서도 기본적인 적대적 생성 신경망을 사용하여 이상 탐지를 진행하였고 Mode Collapse, 실제로 존재하지 않는 데이터 생성 등의 문제를 경험하였다. 7장에서 언급하겠지만, 향후 범주형 이상 시퀀스 데이터를 더욱 효과적으로 분류하기 위해서는 고도화 된 형태의 적대적 생성 신경망을 사용해야 할 것이다.

2.3 적대적 생성 신경망을 활용한 이상 탐지 방법

Thomas et al(2017)의 연구에서 소개된 AnoGAN은 의료 사진의 이상 탐지를 수행하는 분류 모델이다. 합성곱 신경망(Convolution Neural Net)으로 구성하였고 적대적 생성 신경망을 이상 탐지 분야에 본격적으로 활용하게 된 계기가 되었다. 반면, 적대적 생성 신경망을 활용한 시퀀스 데이터 이상 탐지 연구는 이미지나 영상에 비해 연구된 논문이 미비한 실정이다. 물론 Li et al(2018)의 연구에서 순환 신경망의 일종인 LSTM(Long-Short Term Memory)을 기본 적대적 생성 신경망에 접목하여 수치형 시퀀스 데이터의 이상 유무를 분류하는 모형을 제안한 바 있지만, 범주형 시퀀스 데이터에 대해 사용되지 않은 점과 이상 점수(Anomaly Score)을 도출할 때 Salimans et al (2016)이 제안한 자질 매칭 기법 등을 적용하지 않은 점 등은 적대적 신경망을 통한 시퀀스 데이터의 이상 분류에 있어서 시도해야 할 여러 연구 사항이 존재한다는 것을 시사하고 있다. Li가 제안했던 시퀀스 데이터의 이상 탐지를 위한 적대적 생성 신경망 구조도는 <Figure 3>과 같다.



〈Figure 3〉 GAN for Anomaly Detection (Li et al., 2018)

Li가 제안한 모델뿐만 아니라 적대적 생성 신경망의 이상 탐지는 일반적으로 모델을 훈련하는 과정(좌측)과 이상을 탐지하는 과정(우측)으로 분리할 수 있다. 우선 모델 훈련 과정을 살펴보면 학습 시점에서 오토인코더의 이상 탐지 방법과 동일하게 "정상"의 데이터만을 사용하여 분포를 학습하게 된다. LSTM의 기준으로 변형된 목적함수는 (4), (5)와 같다. 여기서 D_{loss}와 G_{loss}는 각각 판별자와 생성자의 손실 값을 의미한다.

$$D_{loss} = \frac{1}{m} * \sum_{i=1}^{m} [log D_{LSTM}(x_i) + log\left(1 - D_{LSTM}\left(G_{LSTM}(z_i)\right)\right)]$$
$$\Leftrightarrow min \frac{1}{\mathrm{m}} * \sum_{i=1}^{m} [-log D_{LSTM}(x_i) - log\left(1 - D_{LSTM}\left(G_{LSTM}(z_i)\right)\right)] \quad (4)$$

$$G_{loss} = \frac{1}{m} * \sum_{i=1}^{m} log\left(1 - D_{LSTM}\left(G_{LSTM}(z_i)\right)\right)$$
$$\Leftrightarrow min \frac{1}{\mathrm{m}} * \sum_{i=1}^{m} log\left(-D_{LSTM}\left(G_{LSTM}(z_i)\right)\right) \quad (5)$$

(5)식의 상단의 $log(1-x)$형태는 0근처의 입력 값에서 0에 가까운 기울기를 가지고 있으므로 표준 정규 분포를 따르는 초기값을 설정하면 학습 초반에 모델 전체의 학습이 느리거나 진행되지 않을 가능성이 있다. 따라서 Li는 학습의 안정성을 위해 (5)식의 하단의 동치 형태로 변형하여 생성자의 최적화를 진행하였다.

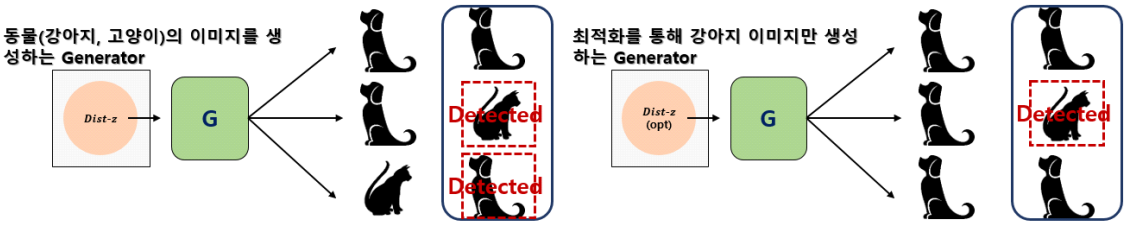
다음으로 <Figure 3>에 이상 탐지 부분에 대해 살펴보자. 적대적 생성 신경망은 정규분포를 가정하는 잠재 변수를 입력 값으로 받아 정상을 모사하는 위조데이터를 생성하고 실제데이터와 위조데이터간의 재건 손실 값 (맨하탄 거리)을 이용하여 이상 점수를 도출한다. 또한, 적대적 생성 신경망은 오토인코더와 다르게 이상 점수를 도출하는 시점에서 판별자로부터 도출되는 판별 손실 값(Discrimination Loss)값을 추가로 이용하는 것을 확인할 수 있다. 여기서 판별 손실 값이란 실제데이터일 확률의 엔트로피 값으로써 정상데이터가 입력되면 충분한 학습 과정을 거친 판별자가 1에 가까운 확률 값을 출력(정상인 데이터로 학습 하였으므로)하고 그 값을 엔트로피에 입력하여 0에 가까운 낮은 손실 값을 도출하도록 한다. 반대로 비정상의 데이터가 입력되면 정상과 상이한 특성의 데이터가 입력되어 판별자는 0에 가까운 확률을 출력하게 되고 이는 높은 엔트로피 손실 값을 갖게 한다.

최종적으로 이상 점수는 재건 손실 값과 판별 손실 값의 가중 합으로써 구하게 된다. 이를 수식으로 나타내면 다음(6)과 같다. 여기서 λ는 가중치, L_G는 재건 손실 값, L_D는 판별 손실 값이다.

$$\mathrm{Score}_{test} = \lambda L_G + (1-\lambda) L_D \quad (6)$$

이상 점수를 구할 때, 선결되어야 할 것은 적대적 생성 신경망에 입력 값인 잠재 변수를 최적화해야 하는 과정이다. 생성 모델의 특성상 어떠한 값이 생성될지 알 수 없기 때문에 탐지 해야 할 데이터와 가장 비슷한 형태의 데이터를 생성하여 이상 탐지를 진행해야 정확한 분류를 할 수 있다. 이에 대한 내용을 <Figure 4>의 예제를 통해 설명하겠다.

<Figure 4>의 예시는 고양이 이미지를 검출해야 하는 적대적 생성 신경망을 나타낸다. 좌측의 모델은 잠재 변수가 최적화 되지 않은 상태이며 우측의 모델은 최적화된 잠재 변수로부터 도출된 결과를 보여준다. 재건 손실 값은 생성자가 도출한 위조 이미지 데이터와 파란색 박스로 강조되어 있는 실제 이미지 데이터의 차를 나타내



〈Figure 4〉 Reason for Latent Variable optimization (example)

므로 생성자가 고양이를 생성하게 된다면 실제 강아지 이미지 데이터의 재건 손실 값이 커지게 되어 고양이를 검출해야하는 상황에서 강아지를 검출하는 잘못된 결과를 야기할 수 있다(좌측의 결과). 즉, 강아지를 검출하기 위해서는 잠재 변수의 최적화를 통해 생성자로부터 강아지 이미지만을 생성하도록 해야 올바른 탐지를 수행할 수 있다(우측의 결과).

2.4 자질 매칭 기법

자질 매칭(Feature Matching)기법이란 위조데이터가 판별자에 입력되었을 때의 은닉층의 벡터와 실제데이터가 판별자에 입력되었을 때의 은닉층의 벡터의 거리를 비교하여 모형에 학습하거나 이상 점수를 도출할 때, 이를 반영하는 방법이다(Salimans et al., 2016). 판별자가 최종적으로 도출하는 확률은 데이터 위조 유무에 관한 이분법적인 정보만을 함의하고 있는 것에 반해, 은닉층에서의 벡터 정보는 실제 데이터임을 입증하는 근거 정보가 함축되어 있어서 다양한 정보를 사용하여 모델을 최적화하거나 이상 점수를 도출할 수 있다. 이를 수식으로 나타내면 (7)과 같다. 여기서 $f(x)$는 실제데이터가 입력된 판별자의 중간층 벡터를 의미하며 $f(g(z))$는 위조데이터의 중간층 벡터를 의미한다.

$$L_D = \sum \left\| E_{x \sim P_{data}} f(x_i) - E_{z \sim P_x(x)} f(g(z)) \right\| \quad (7)$$

3. 데이터 설명 및 전처리

가드몬 시스템이란 사내에서 발생하는 내부 위협을 예방하기 위해 24시간 가동되어 사용자에 의해 발생되는 모든 사건을 로그화하는 시스템을 말한다. 연구에 사용한 데이터는 ㈜H사에서 제공한 가드몬 시스템 사용자 행위 로그 기록이다. 여기서 수집된 데이터는 2주동안(2018. 12. 14 ~ 30) 사용자에 의해 발생된 행위 정보(type)가 포함되어 있고 총 29가지의 범주로 구성되어 있다. 아래의 <Figure 5>는 실제 데이터의 모습을, <Table 1>은 실제 type중 상위 95%를 차지하는 code를 설명한다.

행위는 시간 순서에 따라 데이터에 기록되었지만, 발생한 행위 마다 시간 간격이 동일하지 않은 문제가 있어 일정 시간 동안 발생된 모든 행위를 하나의 시퀀스로 통합하는 방식의 전처리를 수행하였다. 예를 들면 5분 동안 1102라는 Type의 Code가 30번, 1402라는 Type의 Code가 60번 순차적으로 발생하였다면 5분 기준으로 1102가 30개, 1402가 60개가 순차적으로 포함되어 있는 하나의 시퀀스를 만들 수 있다. 이러한

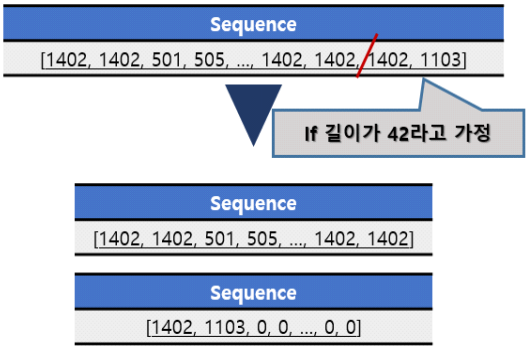
N	Z	AC	AD	AE	AF	AG	AH
gen_date	psptktno	rcv_date	result_code	result_msg	type	user_dev_name	user_id
20181219181452	hamon12_180	2.02E+13			1304	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1302	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1302	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1304	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1304	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1302	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1302	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1304	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1304	DESKTOP-69IGB9<	hamon12_
20181219181452	hamon12_180	2.02E+13			1304	DESKTOP-69IGB9<	hamon12_

〈Figure 5〉 Guard Monitoring user log

〈Table 1〉 Type code (Top 95% occupied)

Type	Description	Type	Description
1402	NET_PEER 생성	1102	프로세스 생성
502	로컬 디스크 파일 수정	1103	프로세스 제거

과정을 거치면 특정 시간 간격마다 시퀀스의 길이가 다른 데이터 셋을 확보할 수 있다. 이후 적대적 생성 신경망에 동일한 길이의 시퀀스를 입력 받을 수 있도록 각 시퀀스를 사용자가 지정한 길이로 분할하고 남은 부분을 제로패딩하여 훈련에 사용될 데이터 셋을 구축하였다. <Figure 6>은 사용자 지정 길이가 40일 때, 분할하는 과정을 보여주는 예이다.



〈Figure 6〉 Sequence chunking process

최종적으로 길이가 통일된 시퀀스 내의 변수들은 각각 29가지 종류의 type과 제로 패딩의 값을 고려한 총 30개의 차원을 가지며 이를 One-Hot 인코딩하였다. 여기서 사용자 지정 길이는 최적의 성능을 도출한 40으로 고정하여 시퀀스를 분할하였다. 시퀀스 길이 별 이상 탐지의 성능 결과는 <Table 2>와 같다.

〈Table 2〉 Performance Difference by Length

Metric	Length 30	Length 40	Length 50
Accuracy	94.62%	98.30%	80.46%
Precision	83.39%	99.80%	98.28%
Recall	91.59%	92.19%	12.10%

40의 길이로 분할한 데이터는 훈련 데이터 52,343개와 탐지에 사용할 테스트 데이터 28,489개로 구성되었고 추후에 진행할 이상 탐지 실험에 사용되었다.

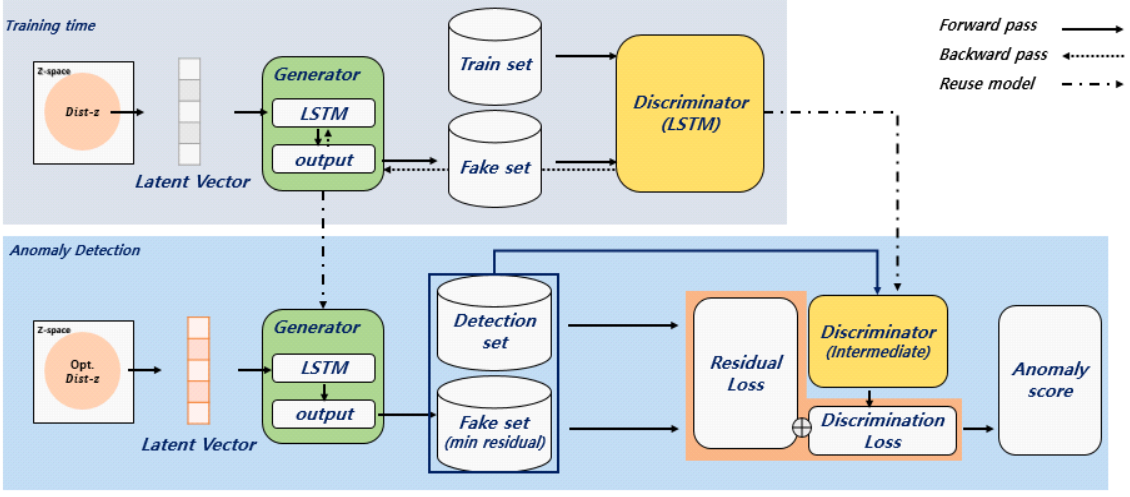
4. 모델 설명

적대적 생성 신경망을 이용한 이상 탐지는 앞서 2장에서 설명한 것과 같이 모델을 학습하는 부분과 학습된 모델을 이용하여 이상을 탐지하는 부분으로 나누어진다는 것을 확인하였다. 본 논문에서의 모델도 동일한 방식으로 모델 설계가 이루어졌다<Figure 7>.

4.1 적대적 생성 신경망의 구조

먼저 입력 받는 잠재 변수의 차원을 어떻게 설정했는지 알아보자. 생성자는 잠재 변수를 우리가 알지 못하는 방식으로 시퀀스 속성과 매핑(Mapping)하기 때문에 사용자가 임의로 차원을 설정할 수 있다. 따라서 차원의 크기를 설정할 때, 실험자가 대상의 속성을 충분히 담을 수 있을 만큼의 잠재 공간의 크기를 사용해야 한다. 하지만 너무 고차원의 변수를 사용하면 잠재 공간상에 분포가 희소해져서 이상 탐지 모델의 성능을 저하할 수 있다. 결론적으로 본 논문에서는 훈련 데이터가 가진 차원보다 낮은 15차원으로 구성된 잠재 변수를 사용하여 잠재 공간상에 데이터의 분포를 적절히 학습할 수 있도록 설계하였다.

이제 <Figure 7>에서 사용하였던 생성자와 판별자의 구조를 살펴보자. 생성자는 각각 32, 64 차원의 2-stacked LSTM의 사용하였다. 그리고 생성자의 출력 부분은 FC(Fully Connected) 층을 활용하여 제로 패딩의 범주와 type범주의 개수를 합한 30차원의 크기로 축소하고 시그모이드 활성 함수를 거쳐 시퀀스 내의 변수가 어떤 type의 범주에 속하는지 확률값을 도출하도록 구성하였다. 판별자는 30차원의 시퀀스를 입력 받아 64차원의 단일 층 LSTM을 거치도록 하였다. 이후, 최종적으로 출력되는 층에서 FC층을 활용하여 1차원의 값을 도출하고 시그모이드 활성 함수를 거쳐 데이터의 위조 유무에 대한 확률 값을 도출하도록 하였다.



〈Figure 7〉 Generative Adversarial Networks for Sequence Anomaly Detection

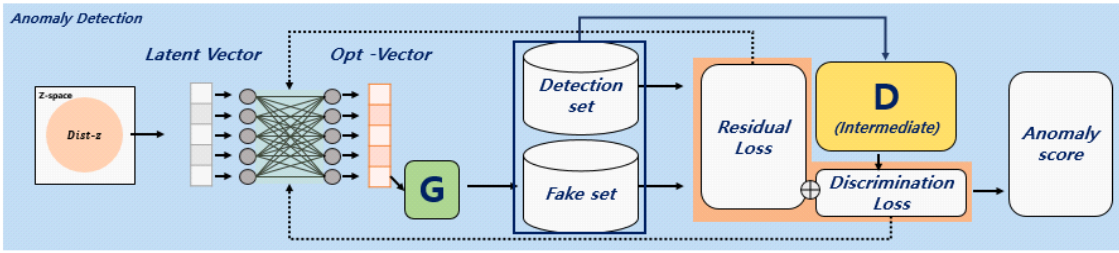
4.2 모델 학습

<Figure 7>의 모델을 학습하는 시점에서 선행 연구와 마찬가지로 “정상”의 데이터만을 사용하였다. 또한, 안정성을 위하여 생성자의 목적함수를 (5)식 하단의 동치 식으로 구성해 최적화를 실시하였다. 학습과정에서의 최적의 학습률(Learnning Rate)은 특정한 범위를 지정하여 그리드 서치(Grid Search)의 방식으로 하이퍼 파라미터를 탐색하였고 생성자 기준으로 0.0001~0.0009, 판별자 기준 0.00001~0.00009의 범위 내에서 수행되었다. 생성자의 탐색 범위를 판별자에 비해 10배 높게 설정한 이유는 판별자의 최적화 속도가 생성자에 비해 빠르다는 것을 이유는 실험을 통해 확인했기 때문이다. 결과적으로 생성자는 0.0002 판별자는 0.00005의 수준에서 Adam 옵티마이저를 활용하였을 때, 최적의 수렴 속도와 학습 안정성이 보장됨을 알아내었다.

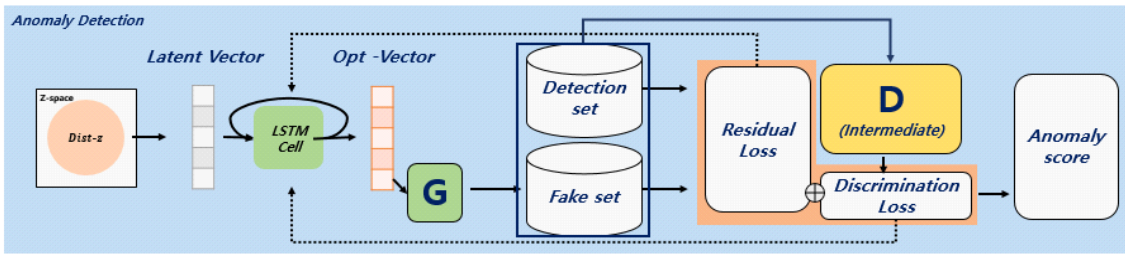
4.3 잠재 변수의 최적화

기존에 연구되었던 적대적 생성 신경망의 이상 탐지 논문(Thomas et al., 2017; Mahdyar et al., 2017; Li et al., 2018)은 잠재 변수를 최적화 할 때, FC층을 활용하여 최적화를 진행하였다. 하지만 기존에 연구에서 최적화 구조에 따른 성능의 변화 실험에 대해서는 어떠한 고찰도 찾아볼 수 없었다. 따라서 본 논문에서는 최적화 구조 변화에 의한 성능 차이를 비교하기 위해, 기존 연구들의 방식과 새롭게 제안한 방식 두 경우로 분리하여 최적화를 진행해 보았다. 먼저 기존의 방식으로 최적화를 진행할 때의 전체 구조는 다음의 <Figure 8>과 같다.

<Figure 8>에서 볼 수 있듯이 잠재 변수를 최적화하는 과정은 적대적 생성 신경망으로부터 도출되는 손실 값을 최소화하는 방식으로 학습이 이루어 진다(<Figure 8>의 점선). 여기서 중요



〈Figure 8〉 Existing latent variable optimizing method



〈Figure 9〉 Modified optimization structure

한 것은 발생한 손실 값으로부터 잠재 변수가 최적화될 때, 기존에 학습이 완료된 생성자(위 그림에서 G)와 판별자(위 그림에서 D)가 과적합(Overfitting) 되지 않도록 추가적인 학습을 진행하지 않는다는 점이다.

본 논문에서는 <Figure 8>과 동일한 방식으로 최적화를 진행하였지만, 분석 도메인의 특성을 고려해서 FC층을 이용한 최적화 방법을 LSTM의 구조로 변경하여 진행하였다. <Figure 9>는 수정된 최적화 방법을 나타낸다.

4.4 이상 점수 도출

적대적 생성 신경망을 활용한 이상 탐지는 훈련이 끝난 생성자와 판별자를 이용하여 탐지데이터에 맞게 잠재 변수를 최적화한 후, 각각 발생되는 재건 손실 값, 판별 손실 값을 가중 합하여 이상 점수를 도출한다. 여기서 본 논문의 차별점은 Li et al(2018)의 연구에서 시도 되지 않았던 자질 매칭 기법을 활용하여 판별자의 중간층을 통해 판별 손실 값을 도출했다는 점이다. 자질 매칭 기법을 활용한 판별 손실 값의 수식은 (6)에서 설명하였다.

한편, 생성자에서 발생한 재건 손실 값은 생성된 위조데이터와 실제데이터의 거리 정보를 가지고 있다. 따라서 생성된 데이터와 실제데이터가 이질적일수록 크기는 커지게 된다. 맨하탄 거리 방식의 척도를 사용하였고 이를 수식으로 나타내면 (8)과 같다.

$$L_G = \sum |x_i - G_{LSTM}(Z_i)| \tag{8}$$

실제 실험에서는 재건 손실 값(8)에 0.9의 가중치와 판별 손실 값(7)에 0.1의 가중치를 주어 이상 점수를 도출하였다. 판별 손실 값을 적게 가중한 이유는 상위 4개의 type에 의해 대부분 구성된 시퀀스에서 의미 있는 패턴 정보가 희소할 것이라고 판단했기 때문이다.

본 연구의 이상 탐지 최종 단계에서는 가중 합으로 산출된 이상 점수를 바탕으로 이상 예측을 수행하기 위해서 분기가 되는 지점의 값(τ)을 연구자가 설정하는 작업이 필요하다. 이를 수식으로 나타내면 (9)와 같고 여기서 Score_{test}는 이상 점수를 의미한다.

$$\mathrm{A}_i^{test} = \begin{cases} 1, if\ \text{Score}_{test} > \tau \\ 0, \quad else \end{cases} \tag{9}$$

따라서 시퀀스 마다 도출된 이상 점수를 가장 낮은 점수부터 가장 높은 점수까지 정렬하고 분기 지점을 변화시키면서 정확도, 민감도, 정밀도가 최상인 지점을 찾는다. 이후, 분기점의 값을 모델에 적용하면 최적의 성능을 도출하는 적대적 생성 신경망의 이상 탐지 모형을 구축할 수 있다.

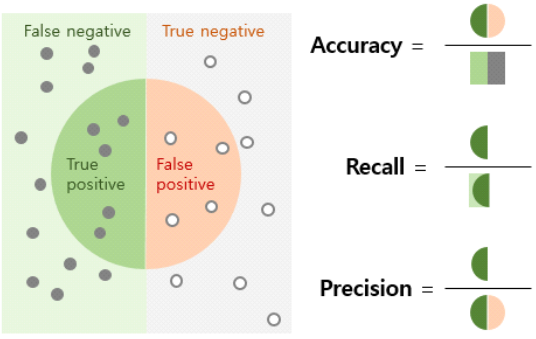
5. 실험 및 결과

실험에 사용한 시퀀스 데이터는 정답이 주어지지 않은 데이터다. 따라서 모델의 성능 비교를 위해서는 “정상”이라고 분류할 수 있는 상황적 가설 설정을 필요로 한다.

크게 두가지의 상황을 기반으로 성능을 측정하였다: 상황1) 상위 95%가 아닌 type이 시퀀스에 포함될 경우 비정상이라 가정. 즉, 드문 행위를 포함하면 비정상. 상황2) 상황1의 경우와 더불어 상위 95%를 차지하는 빈발 행위로 구성된 조합(길이 2)이 드문 패턴일 경우 비정상이라 가정.

5.1 지표 설명

실험에서 사용된 성능 지표는 정확도(Accuracy), 민감도(Recall), 정밀도(Precision)이다. 정확도는 전체 데이터의 개수 중에서 맞춘 개수를 의미하는 지표로 정답 클래스가 편향된 데이터에서는 타당성이 부족한 지표이다. 민감도는 데이터 내에 존재하는 실제 이상 중에서 모델이 맞춘 이상의 개수를 의미하며 이상 탐지 분야에서는 실제 이상을 얼마나 검출했는지 중요하기 때문에 가장 널리 사용되는 지표이다. 정밀도는 모델이 이상이라고 예측한 모든 데이터에서 실제 정확하게 분류한 이상의 수를 의미한다. 이를 이해하기 쉽게 도식화하면 <Figure 10>과 같다.



〈Figure 10〉 Performance metrics

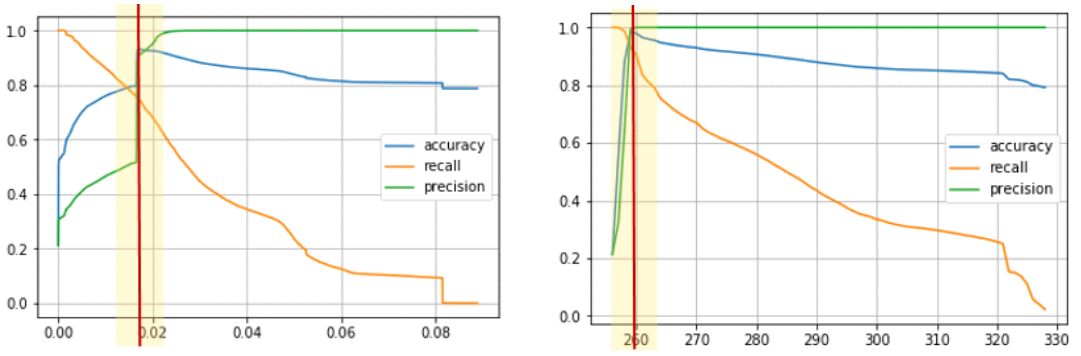
5.2 오토 인코더와의 성능 비교

상황1의 경우에 대해 적대적 생성 신경망과 오토인코더의 성능을 비교해 보았다. 먼저 분기점을 변화시키면서 추적한 정확도, 정밀도, 민감도의 양상을 살펴보자.

<Figure 11>에서 좌측은 오토인코더의 결과이며 우측은 적대적 생성 신경망의 결과를 보여준다. 전체적으로 정확도, 정밀도, 민감도의 변화 추이는 비슷하지만, 적대적 생성 신경망이 더 안정적인 성능을 보이고 있다. 자세한 성능 비교를 위해 <Figure 11>의 각 모델의 최적(적색 선)위치에서 혼동 행렬의 결과를 살펴보도록 하자. 여기서 Epoch과 배치 사이즈는 각각 100과 512로 고정하여 실험하였고 <Table 3>과 <Table4>는 각각 오토인코더와 적대적 생성 신경망의 혼동 행렬 결과를 보여준다.

위 혼동행렬에서 정확도, 민감도, 정밀도를 각각 구해보면 오토인코더의 정확도는 91.04%, 민감도는 75.38%, 정밀도는 81.13%이고 적대적 생성 신경망의 정확도는 98.45%, 민감도는 93.23%, 정밀도는 99.45%이다.

이와 같은 결과가 도출된 경위는 재건 손실 값만을 사용하여 이상 점수를 도출하는 오토인코



〈Figure 11〉 Performance trends of Autoenocoder & GANs

〈Table 3〉 Confusion matrix of Autoencoder

Real \ Prediction	Abnormal	Normal
Abnormal	4,565	1,491
Normal	1,062	21,371

〈Table 4〉 Generative Adversarial Networks of Autoencoder

Real \ Prediction	Abnormal	Normal
Abnormal	5,646	410
Normal	31	22,402

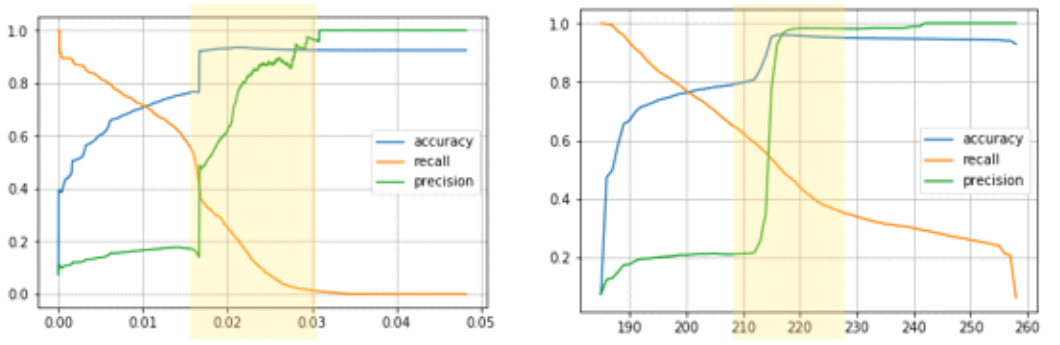
더에 비해 적대적 생성 신경망은 재건 손실 값과 판별 손실 값을 동시에 활용하여 더 많은 정보를 바탕으로 이상 점수를 도출했기 때문이다. 즉, 입력과 출력 형태의 상대적 차이만을 고려하는 오토인코더는 패턴의 정보를 이상 점수에 반영할 수 없어 적대적 생성 신경망에 비해 열등한 결과를 도출한다는 것을 확인하였다.

다음으로 정상의 데이터만을 기반으로 모델을 훈련시키는 기존 방식과 달리 훈련데이터에 “비정상”이라고 가정한 시퀀스가 섞여있을 때, 두 모델의 성능에 얼마나 영향을 받는지 확인하는 강건성(Robustness) 실험을 진행하였다. 적대적 생성 신경망은 분포를 학습하기 때문에 데이터를 직접 인코딩하는 오토인코더에 비해 이상데이터의 영향을 덜 받는다는 것을 가정하고 실험을 통해 이를 증명해 보았다. <Table 5>는 최적의 분기점에서의 강건성 실험 결과를 보여주고 <Figure 12>는 테스트 데이터 상에서 분기점 별 성능 추이를 나타낸다.

<Figure 12>에서 황색 영역을 살펴보면 오토인코더(좌측)은 분기점을 증가시켰을 때, 정밀도의 추세가 적대적 생성 신경망에 비해 단조적으로 증가하는 경향을 보였다. 민감도는 오토인코더가 0으로 급격히 수렴하는 것과는 달리 0.2 이상의 수준을 끝까지 유지하다가 마지막 지점에서 빠르게 감소하는 것을 확인할 수 있었다. 따

〈Table 5〉 Robustness test output

Model	Threshold	Accuracy	Recall	Precision	AUC	Epoch	Batch size	Data type
Autoencoder	0.0167	0.923	0.397	0.480	0.681	100	512	Train
Autoencoder	0.0167	0.923	0.396	0.489	0.681	100	512	Test
GANs	216	0.963	0.535	0.940	0.766	100	512	Train
GANs	216	0.960	0.516	0.928	0.756	100	512	Test



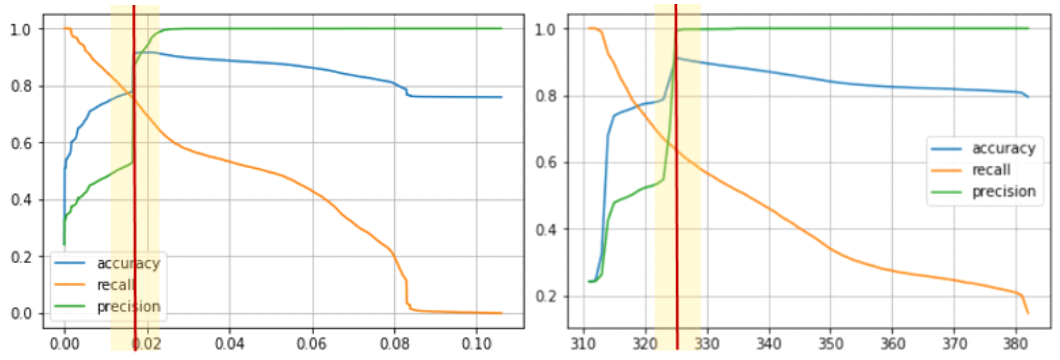
〈Figure 12〉 Model performance trend (Train data with anomaly data)

라서 적대적 생성 신경망은 오토인코더에 비해 이상데이터의 영향을 덜 받는 강건한 모델임이 실험을 통해 입증되었다.

이제 상황2에 대해 오토인코더와 적대적 생성 신경망의 비교 실험을 살펴보도록 하자. 상황2의 실험은 단일 행위로만 이상을 구분하는 것이 아닌 드문 패턴까지 함께 고려하여 이상을 구분할 때, 각 모델의 성능을 확인하는 실험이다.

상황2에서 드문 패턴은 전체데이터의 개수에서 5% 이내의(여기서는 4,401개) 빈도를 가정하였고 (1402, 502), (502,1402)는 전체 데이터에서 각각 928개와 991개의 발생 빈도를 가지고 있어 드문 패턴이라고 정의하였다. 따라서 기존 단일 행위 기준의 정답에서 위 두 패턴을 포함하는 시퀀스를 추가적으로 이상이라고 변경하여 훈련 데이터에서 제외시킨 뒤, 모델을 학습하였다. <Figure 13>은 상황2에서의 성능 변화 추이를 나타낸다.

테스트데이터 기준으로 최적 지점으로 보이는 (적색 선)에서의 성능 지표를 확인해보면 좌측의 오토인코더는 정확도 90.67%, 민감도 73.67%, 정밀도 85.62%를 도출하였고 우측의 적대적 생성 신경망은 정확도 91.14%, 민감도 63.77%, 정밀도 99.18%의 수준의 결과를 도출하였다.



〈Figure 13〉 Performance trend (condition 2)

실험 결과로 확인할 수 있는 시사점은 적대적 생성 신경망이 실험의 모든 경우에서 오토 인코더에 비해 정밀도가 우수했다는 점이다. 하지만 드문 패턴으로 이상이 정의된 데이터 셋으로 학습할 경우, 민감도 측면에서 오토인코더가 더 우수한 성능을 나타내었다. 결과를 분석해보면 분포로부터 위조데이터를 생성하여 도출되는 적대적 생성 신경망의 재건 손실 값은 실제 데이터로부터 직접 도출되는 오토인코더의 재건 손실 값보다 더 일반화된 손실 값이기 때문에 민감도 측면에서 더 안좋은 결과를 도출하였다고 추측할 수 있다.

5.3 잠재 변수의 최적화 구조에 따른 성능 비교

잠재 변수를 최적화할 때, 구조를 FC 레이어를 활용한 것과 LSTM을 활용한 것을 비교해 보았다. <Table 6>은 최적화 구조를 변경하였을 때 테스트 데이터 상의 결과를 보여준다.

정확도 측면에서는 0.1%가량, 민감도는 1% 향상되었지만, 정밀도 차원에서는 0.45% 감소하였다. 이상 탐지 관점에서는 실제 이상인 데이터를 얼마나 잘 검출하는지가 관건이 되므로 민감도의 향상 관점에서 보았을 때, 최적화 구조의 변경은 유의미하다고 볼 수 있다.

5.4 자질 매칭 기법으로 인한 성능 비교

<Table 7>은 이상 점수를 도출할 때, 자질 매칭을 시도하여 성능 평가를 낸 결과이다.

<Table 7>에서의 결과를 살펴보면 자질 매칭 기법이 기존 Li의 실험에 비해 열등한 성능을 도출하였다. 4개의 type으로 전체데이터의 상위 95%를 구성하고 있는 데이터의 특성으로 인해 “이상”이라고 분류되는 시퀀스 내부의 변수도 상위 4개의 type으로 대부분 구성되어 있고 자질 매칭 기법으로 정상과 이상의 패턴의 특성 차이를 파악하기 어려워 기존 모델이 더 나은 성능을 낸 것으로 볼 수 있다. 정리하자면 소수의 type이 전체 데이터의 대부분을 차지할 때, 정상과 이상 시퀀스 간에 유효한 차이를 자질 매칭 기법으로는 획득할 수 없고 확률 엔트로피 값만으로도 정확한 성능을 도출할 수 있음을 시사한다.

〈Table 6〉 Performance comparison according to latent variable optimization structure

Model	Accuracy	Recall	Precision
Fully connected layer	0.9830	0.9219	0.9980
LSTM	0.9843	0.9323	0.9935

〈Table 7〉 Performance comparison according to Feature matching method

Model	Accuracy	Recall	Precision
Feature Matching	0.9843	0.9323	0.9980
Baseline	0.9984	0.9926	0.9998

자질 매칭 기법에서 유효한 결과를 도출하려면 정상의 데이터에서 많은 종류의 type을 포함하고 정상의 type으로 이루어지는 패턴이 다양해야 기존 연구에서의 모델보다 더 향상된 성과를 낼 수 있다는 것을 실험을 통해 미루어 짐작할 수 있다.

6. 결론

본 연구에서는 기존에 시도되지 않았던 범주형 시퀀스 데이터의 이상 탐지를 수행하여 가시적인 성과를 도출하고 이미지나 영상에 치중되었던 적대적 생성 신경망의 연구 활용 범위를 확장할 수 있었다. 이는 쓰임에 있어 범용적으로 사용이 가능하다고 여겨졌던 오토인코더와 동일하게 적대적 생성 신경망에서도 범용적 활용이 가능하다는 것을 시사한다.

본 연구를 통해 확인할 수 있었던 학술적 시사점은 첫번째로 재건 손실 값과 판별 손실 값을 동시에 활용하는 적대적 생성 신경망이 재건 손실 값만을 사용하는 오토인코더에 비해 특수한 상황에서는 민감도가 더 낮을 수 있다는 것이다. 상황1의 실험에 경우, 드문 행위로 정의된 이상에 대해서 정확도, 민감도, 정밀도 모두 적대적 생성 신경망이 우수한 성과를 냈음을 보았다. 반면, 빈발 행위의 조합으로 이루어진 드문 패턴을 함께 탐지하는 상황2의 경우, 비교 모델인 오토인코더보다 민감도 측면에서 성능이 저조하다는 것을 발견하게 되었다. 이는 상위 95%의 발생 행위의 대부분이 4가지 type으로 구성되어있는 데이터의 특수성 때문이다. 드문 패턴으로 정의된 이상을 탐지할 때, 적대적 생성 신경망은 일반화된 재건 손실 값을 사용하기 때문에 정상 패턴의 정보를 포함하는 판별 손실 값에 더 많은 정보가 함축 되어야 올바른 이상 탐지를 수행할 수 있다. 그러므로 정상시퀀스의 패턴 정보와 비정상 시퀀스의 패턴 정보가 상이하지 않는 상황 하에서 판별 손실 값으로부터 획득되는 정보는 모델 전체 성능 향상에 유의한 기여를 하지 못함이 실험을 통해 증명되었다. 이러한 결과를 뒷받침하는 실험은 5.4에 나타나있다. 자질 매칭 기법을 사용하였을 때, 패턴의 정보가 유의했다면 기존 Li et al(2018)에서 제안한 이상 점수를 도출하는 방식보다 향상된 성능을 도출했을 것이다. 따라서 소수의 행위가 데이터의 대부분을 차지하는 경우, 드문 패턴 정보를 반영하는 이상을 검출할 때에는 민감도 측면에서 오토인코더를 활용하는 것이 좋고 드문 행위 정보만을 가정하여 이상을 검출할 때는 적대적 생성 신경망이 사용되어야 한다.

두번째는 비정상데이터가 혼재된 훈련데이터를 사용하였을 때, 적대적 생성 신경망의 성능 감소 폭이 오토인코더에 비해 작다는 것을 실험을 통해 입증했다는 것이다. 본 연구에서는 정상과 이상을 특정한 상황을 가정하여 실험을 진행하였지만, 현실의 이상 탐지 데이터는 정답이 주어지지 않는 데이터를 사용하므로 모델의 훈련 과정에서 실제 비정상데이터가 훈련데이터에 포함될 수 있다. 따라서 적대적 생성 신경망은 첫번째 시사점에서 언급된 것과 같이 특수한 상황에서 일부 저조한 성능을 낼 수 있겠지만, 비정상이 혼재된 데이터를 기반으로 훈련했음에도 성능 변화가 오토인코더에 비해 적다는 것을 증명하여 실제 산업에서 사용될 경우 활용가치가 더 높다고 말할 수 있겠다.

세번째는 잠재 변수를 어떠한 방식으로 최적화하는지에 따라 모델 성능이 의존적이라는 사

실을 입증한 것이다. 현존하는 적대적 생성 신경망의 이상 탐지 논문을 살펴보면 이에 대한 구체적인 연구를 찾아보기 어렵다. 따라서 본 논문에서 분석 도메인에 맞는 최적화가 성능에 긍정적인 변화를 발생시키는 것을 증명하였으며 향후 연구자들이 이에 대한 추가적인 연구가 수행되어야 함을 보여주었다.

7. 향후 연구

본 논문의 한계점은 적대적 생성 신경망으로부터 도출되는 이상 점수가 전반적으로 높다는 것이다. 오토 인코더의 최적의 분기점을 살펴보면 0.02의 수준 안에서 모델의 최적 성능이 도출되지만, 적대적 생성 신경망은 최적의 분기점이 200에서 300 초반에 걸쳐 다양하게 형성되어 있다. 이는 정상과 이상을 분류하는 성능이 적대적 생성 신경망이 상대적으로 우수하나 생성되는 위조 데이터의 품질이 오토인코더에 비해 열등하다는 것을 반증하는 것이다.

생성되는 데이터의 품질이 좋지 않은 이유를 분석해보면 현재의 모델은 잠재 변수가 연속형의 정규분포를 가정하여 연구를 진행하였기 때문이다. 범주형 도메인에서 실제로 존재하지 않는 범주 사이의 데이터를 생성하고 범주 사이의 값들은 현재 연구기준으로 모두 손실 값이 되어 높은 이상 점수가 도출된다. 따라서 향후에는 이를 해결하기 위한 범주형 형태에 특화된 적대적 생성 신경망의 활용이 필요할 것으로 예상된다.

참고문헌(References)

Sun, B., P. B. Luh, Q. S. Jia, Z. O'Neill, and F. Song, "Building energy doctors: An spc and kalman filter-based method for system-level fault detection in hvac systems", *IEEE Transactions on Automation Science and Engineering*, Vol.11, No.1, (2014), 215~229.

Du, Z., B. Fan, X. Jin and J. Chi, "Fault detection and diagnosis for buildings and hvac systems using combined neural networks and subtractive clustering analysis", *building and environment*, Vol.73 (2014), 1~11

Koturwar, P., D. Mukhopadhyay and S. Griase, *A survey of classification techniques in the area of big data*, Department of Information Technology Maharashtra Instititute of Techonology, 2014, Available at https://arxiv.org/abs/1503.07477 (Downloaded 10 June, 2019)

Pimentel, A.F M., D. A. Clifton, L. Clifton and L. Tarassenko, "A review of novelty detection", *Signal Processing*, Vol.99, (2014), 215~249

Ye, N., S. Vilbert and Q. Chen, "Computer intrusion detection through ewma for autocorrelated and uncorrelated data", *IEEE transactions on reliability*, Vol.52, No.1, (2003).

He, X., Z. Wang, Y. Liu, and D. H. Zhou, "Least-squares fault detection and diagnosis for networked sensing systems using a direct state estimation approach", *IEEE Transactions on Industrial Informatics,* Vol.9, No.3, (2013), 1670~1679.

Ye, N. and Q. Chen, "An anomaly detection technique based on a chisquare statistic for detecting intrusions into information systems",

Quality and Reliability Engineering International, Vol.17, No.2, (2001), 105~112.

Dai, X. and Z. Gao, "From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis", *IEEE Transactions on Industrial Informatics*, Vol. 9, No. 4, (2013), 2226~2238.

Goh, J., S. Adepu, M. Tan and Z. S. Lee, *Anomaly Deetection in cyber physical systems using recurrent neural networks*, IEEE, Sigarpore, 2017.

Esteban, C., S. L. Hyland and G. Ratsch, *Real-valued (medical) time series generation with recurrent conditional gans*, Tri-Institutional Training Program in Computational Biology and Medicine Weill Cornell Medical, 2017. Available at https://arxiv.org/abs/1706.02633 (Downloaded 13 June, 2019)

Zenati, H., C. S. Foo, B. Lecouat, G. Manek and V.R Chandrasekhar, *Efficient gan-based anomaly detection*, ICDM, 2018. Available at https://arxiv.org/abs/1802.06222 (Downloaded 1 May 2019)

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets", *Advances in neural information processing systems*, Vol. ACM, (2014)

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans", *In Advances in Neural Information Processing Systems*, Vol.29, (2016), 2226~2234.

Hinton, G. E and R. R. salakhutdinov, "Reducing the dimensionality of Data with Neural Network", *Science*, Vol.313, No. 5786, (2006), 504~507.

Fraink, J., *Artificial intelligence and intrusion detection: Current and future directions*, Division of Computer Science, University of California, 1994. Available at https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.5769&rank=1

Iwan, S., A. Prugel-Bennett and G. Wills, *Networked digital Technologies*, Springer, Dubai, 2012.

Deecke, L., R. Vandermeulen, L. Ruff, S. Mandt and M. Kloft, *Anomaly Detection with Generative Adversarial Networks*, 2018. Available at https://openreview.net/forum?id=S1EfylZ0Z (Downloaded 13 June 2019)

Sakurada, M. and T. Yairi, *Anomaly detection using autoencoders with nonlinear dimensionality reduction*, Machine Learning for Sensory Data Analysis, Dunedin, 2014.

Li, D., D. Chen, J. Goh and S-K. Ng, *Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, DBLP, London, 2018.

Schlegl T., P. Seeb¨ock, S. M. Waldstein, U. Schmidt and G. Langs, *Computer vision and pattern Recognition*, IPMI, North Carolina, 2017.

Abstract

Anomaly Detection for User Action with Generative Adversarial Networks

Choi, Nam woong* · Wooju Kim**

At one time, the anomaly detection sector dominated the method of determining whether there was an abnormality based on the statistics derived from specific data. This methodology was possible because the dimension of the data was simple in the past, so the classical statistical method could work effectively. However, as the characteristics of data have changed complexly in the era of big data, it has become more difficult to accurately analyze and predict the data that occurs throughout the industry in the conventional way. Therefore, SVM and Decision Tree based supervised learning algorithms were used.

However, there is peculiarity that supervised learning based model can only accurately predict the test data, when the number of classes is equal to the number of normal classes and most of the data generated in the industry has unbalanced data class. Therefore, the predicted results are not always valid when supervised learning model is applied. In order to overcome these drawbacks, many studies now use the unsupervised learning-based model that is not influenced by class distribution, such as autoencoder or generative adversarial networks.

In this paper, we propose a method to detect anomalies using generative adversarial networks. AnoGAN, introduced in the study of Thomas et al (2017), is a classification model that performs abnormal detection of medical images. It was composed of a Convolution Neural Net and was used in the field of detection. On the other hand, sequencing data abnormality detection using generative adversarial network is a lack of research papers compared to image data. Of course, in Li et al (2018), a study by Li et al (LSTM), a type of recurrent neural network, has proposed a model to classify the abnormities of numerical sequence data, but it has not been used for categorical sequence data, as well as feature matching method applied by salans et al.(2016). So it suggests that there are a number of studies to be tried on in the ideal classification of sequence data through a generative adversarial Network. In order to learn the sequence

* Department of Industrial Engineering, Yonsei University
** Corresponding Author: Wooju Kim
Graduate School of Industrial Engineering, Yonsei University
50 Yonsei-ro Seodaemun-gu, Seoul, Korea
Tel: +82-2-123-7754, Fax: +82-2-123-7754, E-mail: wkim@yonsei.ac.kr

data, the structure of the generative adversarial networks is composed of LSTM, and the 2 stacked-LSTM of the generator is composed of 32-dim hidden unit layers and 64-dim hidden unit layers. The LSTM of the discriminator consists of 64-dim hidden unit layer were used.

In the process of deriving abnormal scores from existing paper of Anomaly Detection for Sequence data, entropy values of probability of actual data are used in the process of deriving abnormal scores. but in this paper, as mentioned earlier, abnormal scores have been derived by using feature matching techniques. In addition, the process of optimizing latent variables was designed with LSTM to improve model performance. The modified form of generative adversarial model was more accurate in all experiments than the autoencoder in terms of precision and was approximately 7% higher in accuracy.

In terms of Robustness, Generative adversarial networks also performed better than autoencoder. Because generative adversarial networks can learn data distribution from real categorical sequence data, Unaffected by a single normal data. But autoencoder is not. Result of Robustness test showed that he accuracy of the autocoder was 92%, the accuracy of the hostile neural network was 96%, and in terms of sensitivity, the autocoder was 40% and the hostile neural network was 51%.

In this paper, experiments have also been conducted to show how much performance changes due to differences in the optimization structure of potential variables. As a result, the level of 1% was improved in terms of sensitivity. These results suggest that it presented a new perspective on optimizing latent variable that were relatively insignificant.

Key Words : Autoencoder, Anomaly Score, Feature matching, Generative Adversarial Nets-Anomaly Detection, Optimizing latent variable

Received : June 4, 2019 Revised : July 22, 2019 Accepted : August 2, 2019
Publication Type : Conference(Fast-track) Corresponding Author : Wooju Kim

저 자 소 개

최남웅
연세대학교 산업공학과에서 석사과정 재학 중이다. 주요 관심 분야는 자연어 처리, 이상 탐지, 빅데이터 분석 등이다.

김우주
1987년 연세대학교 BBA 과정 학사 학위를 취득하고, 1994년 KAIST 경영과학 박사를 취득하였으며, 현재 연세대학교 정보산업공학과 교수로 재직 중이다. 관심분야는 시맨틱 웹, 시맨틱 웹 환경의 의사결정지원 시스템, 시맨틱 웹 마이닝, 지식관리 및 인공지능 웹 서비스이다.