|  |  |  |  |
| --- | --- | --- | --- |
| **Unsupervised Anomaly Detection**  **For Time-Series data**  제출자 : 숭실대학교 AI 데이터 연구단 박준혁 |  |  |  |

**Mother Paper**

텍스트이(가) 표시된 사진

자동 생성된 설명

**0. Abstract**

**FPR(False Poitive Rate)을 최소화하고** 보다 정확한 머신 러닝 모델을 구축하기 위해서, 정상 행동에서 비정상 변화점을 발견하고, 이를 이용하여 비정상 데이터인, 이상치를 구별하는 일은 매우 중요한 과정이다. 본 논문에서는 **다변량 Time-Series 데이터에서 변화점 탐지**를 수행하여 이상치 탐지 성능을 높이는, 규칙 기반 결정 시스템을 제안한다. 최종적으로 기존의 비지도 학습 모델과 딥 러닝 모델을 사용하여 5개의 변경 점 탐지 알고리즘을 구축하였다. 기재된 실험 결과는 우리가 제안한 규칙의 정확성을 검증할 뿐만 아니라 변화점을 제거하여 이상 탐지를 개선할 수 있다는 가정을 강화한다.

이상치 탐지는 시계열 데이터에서 예측과 예상의 질을 향상시키기 위해 중요한 부분이다. 우리의 구조는 자동적으로 실제 이상치를 발견하고, 변화점을 활용하여 FPR을 제거하는 파이프라인을 사용한다. 추가적으로 우리는 시계열 데이터에 변화가 시작되는 지점에 기초한 새로운 평가 지표를 제안한다 : **low MSE**

**1. Introduction**

우리는 물 자원 관리와 분배 네트워크의 관련된 시계열 데이터 분석에 대해 집중하고자 한다. 물 소비의 감소는 다음과 같은 여러 변수를 고려하여, 비수익과 수요를 최소화하기 위해 물 분배망의 개선이 필요한 문제이다. **1) 현재 물 분배망의 상태, 2) 변화 수요, 3) 필요한 수요량**

물 자원 관리 성능을 개선하고, 물 소비와 손실을 최소화하기 위해선, 영리하고 자동화된 의사 결정 시스템이 필요하다. 그것은 ML, DL을 사용하여 실시간 자원 변화를 분석하고, 사용자에게 물 분배망에서 발견된 이상치를 알려주어야 한다. 추가로, 우리가 제안한 시스템은 사전, 사후 정비를 가능하게 하는 효과적인 절차를 채택하기 위해 최종 사용자에게 물 분배망의 변경 사항을 실시간으로 알린다. **사전 사후 정비 솔루션**은 물 분배망에서 결함이나 장비 고장이 일어날 경우 발생하는 자원 손실 비용을 경감하기 위한 방법이다.

이상 탐지 기반 접근 방식은 잠재적으로 더 많은 공격을 탐지할 수 있지만, 현재 시스템은 높은 거짓 양성율로 인해 어려움을 겪고 있다. 이상치 탐지에 있어서, 변화점 탐지는 시계열 데이터에서 중요한 역할을 한다.

현재 대다수의 이상치 탐지 시스템은 생성된 데이터에 대하여 해석을 위한 사람들의 노력이 필요하다. 또한, 위 과정은 상호 연결된 디바이스의 수가 기하급수적으로 증가하며 점점 더 복잡해지는 양상이다. 게다가, **많은 이상치 탐지 모델 알고리즘은 변화점을 이상치로 분류하는 경향이 있기 때문에 높은 거짓 양성율을 갖는다.** 따라서 이상치와 변화점을 매칭하는 과정을 자동화함으로써, 몇 가지 특이 케이스에 한해서만 사람의 해석 과정이 필요한 시스템을 구축하게 된다. 위 절차는 이상치와 변화점을 높은 향상도로 구별하는 규칙 기반 결정 시스템을 통해 개선될 수 있다.

우리는 현재 문헌에서 제시된 기존 이상치 탐지 방법들의 단점을 고려하여 아래와 같은 문제에 대한 해답을 구하고자 한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

위 네 문제에 대한 해답을 구하기 위해, 우리는 IoT 시계열 데이터에서 변화점 탐지를 통해 이상치 탐지 성능을 높이고자 한다. 우리가 사용한 **5개의 이상치 탐지 알고리즘은 Gaussian Distribution(GD), K-Means, Isolation Forest(IF), OC-SVM(One-Class Support Vector Machine) 그리고 Autoencoders(AEs)**이다. **그리고 변화점 탐지를 위해 사용한 또다른 5개의 알고리즘은 Window-based segmentation(WinSeg), Binary Segmentation(BinSeg), Bottom-up Segmentation(BottomUp), Pruned Exact Linear Time(PELT) and Exact Segmentation dynamic Programming Model(OPT)이다**. 또한, 우리는 변화점의 지지도와 이상치의 지지도에 근거한 새로운 평가 지표를 제안한다.

**2. Related Work**

‘Kieu’는 **Sparsely-connected Recurrent AutoEncoder ensembles**에 기반한 시계열데이터의 이상치 탐지 알고리즘을 제시하였다. 그들은 다수의 AE를 결합함으로써, 이러한 앙상블이 일부 AE의 과대적합을 줄이는 경향이 있기 때문에 전체적인 탐지 정확도가 향상된다고 주장하였다.

RNN은 시계열 데이터에서 변화점과 이상치 탐지에서 성공적인 성능을 보였다. ‘Saurav’는 정상 행동상의 갑작스럽거나 정기적인 변화를 고려하여 시계열 이상치 탐지를 위한 RNN 기반 시간적 모델을 제안하였다. **기존의 RNN 모델은 사라지는 기울기 문제를 가지고 있다**, 그러므로, 새로운 RNN 모델이 설계되었는데, 바로 **LSTM(Long Short-Term Memory)이다.** LSTM 기반 솔루션은 시계열 분석 중에서도 특히 이상치 탐지 부분에서 널리 사용되었다. 제안된 솔루션은 **Sliding Window**를 사용하여 알고리즘의 매개 변수를 반복적으로 업데이트함으로써 온라인 탐지를 제공한다.

‘Munir’은 비 주기 스트리밍 데이터에서 주기성 및 계절성 관련 포인트 이상치를 감지하는데 있어서 CNN의 효과에 대해 다루었다. **제안된 CNN 기반 솔루션, ‘DEEPANT’은 시계열 예측기와 이상치 탐지기 이와 같이 두 개의 모듈로 구성되어 있다.** 각 예측 값은 이상치 탐지기를 거쳐 정상인지 아닌지 태그를 달게 된다. ‘DEEPANT’ 실험의 중요한 결론은 비록 LSTM 기반 솔루션이 LSTM의 장기 추세 추출 능력으로 인해 시계열 데이터에 대해 좋은 성능을 보이지만, 매개 변수 효율면에서 CNN 또한 좋은 대안이라는 것이다.

몇몇 **저자들은 데이터의 전처리 과정이 매우 중요하며**, 탐지 성능에 중요한 영향을 미칠 것이라고 주장한다. ‘Zhang’은 **전처리, 예측, 이상치 탐지 이상 세 개의 하이브리드 모델로 구성된 솔루션을 제안하였다**. 전처리 단계에서는 입력 데이터를 분석하고, 분명한 이상치들을 제거한다. 그리고 두 번째 단계에는 전처리된 데이터를 새로운 기능 및 고정 내장 모드로 차원 축소하고, 예측 값을 얻기 위해 분류, 회귀, 자기회귀 등을 적용한다. 또한, 예측값으로부터 이상치를 탐지하기 위해 지수 가중 이동 평균 방법을 사용하였다. 위 실험을 통해 얻은 결과를 통해 저자는 제안된 모델의 앞선 두 단계가 모델의 정확도와 Robustness를 향상시킨다는 것을 발견하였다.

현재 문헌을 분석해본 결과, 기존 솔루션들은 흥미로운 여러 파이프라인을 제안했지만, 대부분 대규모의 실시간 시계열 데이터로 철저한 시험을 수행하지는 않았다. 또한, 우리가 아는 한**, 잘못된 양성 비율을 낮추기 위해 변화점 알고리즘을 사용하는 이상 탐지 솔루션은 현재 없다는 것이다.**

**3. Methodology**

**3.2 Time Series Outlier Detection**

시계열 데이터에 대한 정의와 더불어, 우리의 시스템 구조에 사용될 **이상치와 변화점을 연관 짓는** 몇 가지 모델에 대해 의논해보자.

특정 기간 동안 기록된 순차적 데이터를 설명하는 통계적 수학적 모델을 시계열이라고 한다. 일변량 시계열은 하나의 시간 도메인 변수를 의미하고, 다변량 시계열은 최소 2개 이상의 시간 도메인 변수를 가지는 것을 의미한다. 위에서 말한 각 변수들은 각 변수의 이전 값과 다른 변수들에 의존되어 결정된다.

텍스트이(가) 표시된 사진

자동 생성된 설명 위 구성요소들은 시계열 데이터를 차원 축소 하는데 이용된다. (xt = mt + st+ yt)

특이치 탐지 문제는 특이치 점수 st를 사용하여 시계열 X의 데이터 점 xt의 순위를 매기는 것을 의미하며, st 값이 높을수록 특이치가 될 가능성이 높다.

3.2.1 Gaussian Distribution

GD or 정규분포는 확률적 통계 접근법으로 이상치 탐지에 사용된다. 데이터 포인트가 정규분포를 따르는지 결정하기 위해, PDF(Probability Density Function)가 사용된다. 아래 식은 GD의 PDF의 일반적인 형태를 보여준다.

텍스트, 시계, 손목시계, 게이지이(가) 표시된 사진

자동 생성된 설명

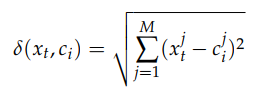
GD의 경우, 이상치는 평균가 멀리 떨어진 지점의 데이터이다. 해당 포인트는 평균으로부터 큰 표준 편차를 가진다. 거리를 탐지하는 한 가지 방법은 타원을 사용하여 데이터 점의 중심을 잡아 데이터에 대한 공분산 추정치를 구하는 것이다. 마할라노비스 거리는 데이터 포인트가 이상치 인지, 그리고 타원을 결정할 때 추정치로 쓰인다. 타원 경계 밖에 모든 지점은 이상치로 고려된다. 아래 수식은 두 지점 간의 마할라노비스 거리를 나타낸다.

텍스트, 손목시계이(가) 표시된 사진

자동 생성된 설명

3.2.2 K-Means

K-Means는 k의 센트로이드 포인트 업데이트와 클러스터 할당을 반복하는 알고리즘이다. K-Means의 정지 기준은 클러스터의 센트로이드가 더 이상 변하지 않거나 정해진 반복 횟수를 끝마쳤을 때이다.



3.2.3 Isolation Forest

IF는 RF에 기반한 비지도 이상치 탐지 알고리즘이다. IF는 서브 샘플을 임의로 선택된 특성과 값의 이해 분할한다. 데이터를 분할하면 데이터를 더 잘 나타내는 속성 값이 초기 분할에서 분리될 가능성이 높기 때문에 이상치에 대한 경로가 더 짧아지는 트리가 생성된다. 이상치가 IF의 루트에 가까울수록 점수는 경로 길이의 함수이다.

3.2.4 One-Class Support Vector Machine

OC-SVM은 한 클래스의 요소를 포함하는 데이터 셋에 대해 훈련된 비지도 기계 학습 알고리즘이다(그 클래스가 ‘정상’ 데이터). OC-SVM은 전체 데이터셋에 대한 지지도를 추정하는 함수를 모델링한다. 위 접근은 데이터셋 내 대표 지점을 탐지하고, 이상치를 고립시키기 위해 사용된다. 아래 함수는 OC-SVM 함수이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

3.2.5 AutoEncoders

AE는 입력 시계열을 출력 벡터로 재구성하는 신경망 모델이다. 본 신경망의 구조는 Encoder와 Decoder 두 개의 구성요소로 이루어진다. 목표는 각 데이터 포인트의 재구성 오차를 최소화하는 것이다.

**3.3 Time Series Change Point Detection**

시계열 데이터의 변화점 검출 문제는 시계열의 속성(예: 평균, 분산 등)이 갑자기 변하는 시점을 결정하는 문제를 다룬다. 변화점 탐지 방법은 변화점 지점의 수 : K에 대한 사전 지식에 따라 달라진다. 변화점 탐지 문제는 시계열 데이터에 대한 최적의 분할 방법을 찾는 것으로 구성된다. 최고의 분할은 비용 함수를 최소화하는 모델이다.

3.3.1 Window-Based Segmentation Model (WinSeg)

WinSeg 알고리즘은 변화점을 발견하기 위해 두 개의 인접한 창 간의 불일치도를 계산한다. 두 개의 창은 시계열 데이터 X의 지점을 따라 이동하는 창이다. 불일치는 바로 과거, 즉 왼쪽 창과 가까운 미래, 즉 오른쪽 창 사이의 시간 지수에 대해 계산된 값이다. 위 방법을 사용하여 알고리즘은 고점 즉, 더 충이 서로 다른 구간을 지나갈 때 가장 큰 값을 계산하고, 불일치 곡선을 계산한다. 변화점은 곡선에서 최고점 탐지 절차를 걸쳐 결정된다.

3.3.2 Binary Segmentation Model (BinSeg)

BinSeg는 비용합을 최소화하는 변화점을 탐색하는 탐욕적 순서 알고리즘이다. 시계열 데이터는 첫 변화점에서 두 개로 분할되고, 새로운 서브셋에서 계산을 다시 수행한다. 알고리즘은 필요한 변화점의 수가 충족되거나 정지규칙을 만족하며 멈춘다.

3.3.3 Bottom-up Segmentation Model (BottomUp)

BottomUp 분할은 많은 변화점으로 시작하여 유의하지 않은 것을 연속적으로 삭제하는 빠른 신호 분할을 하는데 사용되는 순차적 접근법이다. 따라서 시계열은 정규 그리드를 따라 많은 세그먼트로 분할된 다음, 인접한 세그먼트가 얼마나 유사한지에 따라 연속적으로 병합된다.

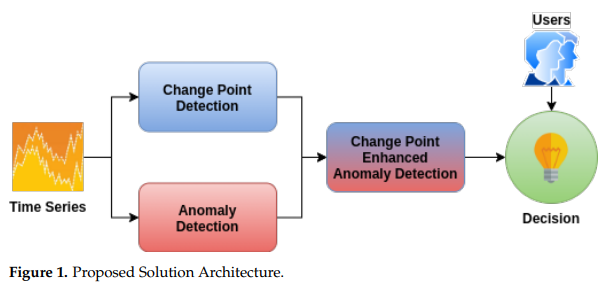
3.3.4 Pruned Exact Linear Time Model (PELT)

변경점의 수를 알 수 없는 경우, 벌점 함수를 최소화하기 위해 정확한 분할을 사용하여 최적의 솔루션을 계산할 수 있다. PELT 모델은 벌점이 선형적일 때 정확한 솔루션을 찾을 수 있다. 명시된 가지치기 규칙은 각 샘플이 잠재적 변경점 인지를 테스트한다.

3.3.5 Exact Segmentation Dynamic Programming Model (OPT)

변경점의 수를 알고 있을 때, 동적 프로그래밍 모델을 활용한 정확한 분할을 통해 최적의 솔루션을 계산할 수 있다. OPT 알고리즘은 목적 함수의 가변성을 사용하여 재귀적으로 문제를 해결한다. 이러한 관찰에 따르면 Xt의 모든 하위 시리즈의 K-1번째 요소를 가진 최적의 분할을 알 수 있으므로, 최적의 분할의 변화 지점을 알 수 있다. 위 과정을 반복하여 모든 세그먼트들을 계산한다.

**4. Proposed Solution**



우리가 제안한 솔루션의 구조는 4 개의 모듈을 포함하고 있다. 이상치 탐지 모듈은 입력 시계열에서 이상치 탐지를 위해 전문화된 비지도 머신러닝 알고리즘을 사용한다. 그리고 변경점 탐지 모듈은 같은 입력 시계열에서 변경점 탐지를 위해 다른 기술을 사용한다. 변경점이 증강된 이상치 탐지 모듈은 이전의 두 개의 모듈의 결과를 비교하여 출력을 만들고, 각 이상치에서 신뢰도 점수를 계산한다. 마지막으로 위 결과는 정해진 규칙에 의거한 결정 모듈을 통해 분석된다.

**4.1 Anomaly Detection Module**

이상치 탐지 모듈은 파이썬 내 최신 사이킷런 머신 러닝 패키지를 사용하여 이루어진다.

Gaussian Distribution : 우리의 구현 상에서는 타원의 원주를 GD의 공분산으로 사용하였다. 이상치의 비율은 0.01로 초기화하였다.

K-Means : 초기화 알고리즘으로 K-MEANS++을 사용하였고, 차원의 저주를 방지하고 차원을 축소하고자 PCA를 사용하였다. 그리고 최적의 클러스터 수를 얻기 위해 Elbow Method를 사용함. 가장 가까운 중심으로부터의 거리가 주어진 임계값을 넘어서는 점은 이상치로 간주된다.

Isolate Forest : 100개의 IF 모델을 붓스트랩 방식을 추출된 서브 샘플을 통해서 학습하였다. 각 시계열에 대해 이상치의 비율은 교차 검증을 사용하여 결정해야 한다.

One-Class Support Vector Machine : 커널 함수를 RTF로 초기화한다. 그리고 교차 검증을 사용하여, 학습 에러의 상한선을 0.01로 설정한다.

Autoencoder : 13개의 완전 연결 은닉층을 가지는 신경망 모델을 구축하였다. 각 은닉층의 퍼셉트론 수는 [M, 64, 32, 16, 8, 4, 2, 4, 8, 16, 32, 64, M]이다.

**4.2 Change Point Detection Module**

모든 채택된 변경점 탐지 알고리즘에 대해, 우리는 동일한 매개변수 구성을 사용하였다 : 모델은 l2-norm을 사용하여 초기화되며 두 변경점 사이의 최소 세그먼트와 길이는 2로 설정된다. PELT는 자동으로 정지하기 전에 최적의 포인트 수를 결정한다. 다른 알고리즘에 대해선 PELT로 계산된 값을 최적의 중단점으로 설정한다.

**4.3 Change Point Enhanced Anomaly Detection Module**

변경점 증강 이상치 탐지 모듈은 데이터 포인트가 이상치가 될 확률을 계산한다. 데이터 포인트가 각각 이상치 혹은 변경점이 될 지지도를 계산해서 향상도를 결정한다.

데이터 포인트가 이상점과 변경점으로 모두 분류될 때, 신뢰도 함수는 변화점에 의해 유입된 명백한 잘못된 긍정을 제거할 수 있다. 최종 결정을 인간 운영자에게 맡기는 경우, 이 기능은 정보에 입각한 결정을 내리는 데 도움이 된다.

텍스트이(가) 표시된 사진

자동 생성된 설명

위 함수는 또한 효과적인 사전,사후 정비를 가능케 한다. 신뢰도 함수의 값을 평가함으로써 시스템의 올바르게 작동하는지 판단하고, 적시에 사전 정비를 달성할 수 있다. 그러므로, 우리가 이상치를 변경점으로 판단한다면, 그것은 시스템이 정상적으로 작동한다는 것이다. 반면에 오직 이상치로만 판별한다면. 우리는 높은 확신을 가지고 그 시스템에 결함이 있다고 가정할 수 있다. 사후 정비는 근본적인 실패의 이유를 감독하고 수정하는데 집중한다.

※ 이상치로만 판단될 경우 : 이상치

변경점으로만 판단될 경우 : 정상

이상치와 변경점이 비슷한 경우 : 사람의 해석

**4.4 Decision Module**

의사 결정 모듈은 규칙을 사용하여 사람 관리자에게 경고를 보낼지 혹은 시스템이 자동으로 결정을 내릴지 결정한다. 대부분의 경우는 시스템에 의해 해석되기 쉽고, 자동 응답이 주어진다. 드문 경우에 한에서 사람의 해석이 필요하다. 그러므로, 시스템은 관리자에게 경고하고, 정보에 근거한 결정을 내리도록 할 필요가 있다.

시스템은 아래와 같은 빈발 케이스에 대해 자동적으로 결정을 내릴 수 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그리고 아래와 같이 드문 경우에는 여전히 사람의 해석이 필요하다.

텍스트이(가) 표시된 사진

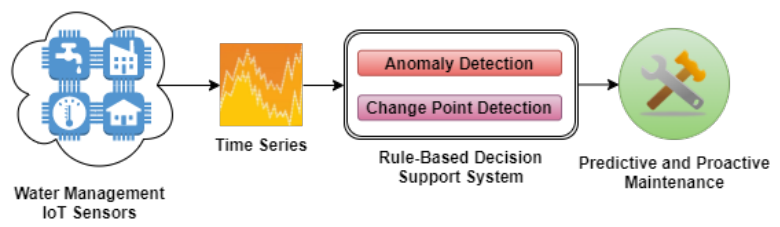
자동 생성된 설명

H1, H2, 그리고 H3에 대해서는 데이터 포인트가 이상치 인지 변경점 인지에 관계없이, 사람 관리자가 데이터 포인트에 대해 분석하고 결정을 내려야 한다. 유사 경우에 기조하여, 우리는 자동으로 문제를 해결하고 사람 관리자에게 경고를 해줄 아래 표와 같은 결정 규칙을 제안한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

**※ confidence(x)가 0.5에 가까우면 사람의 해석이 필요하다.**



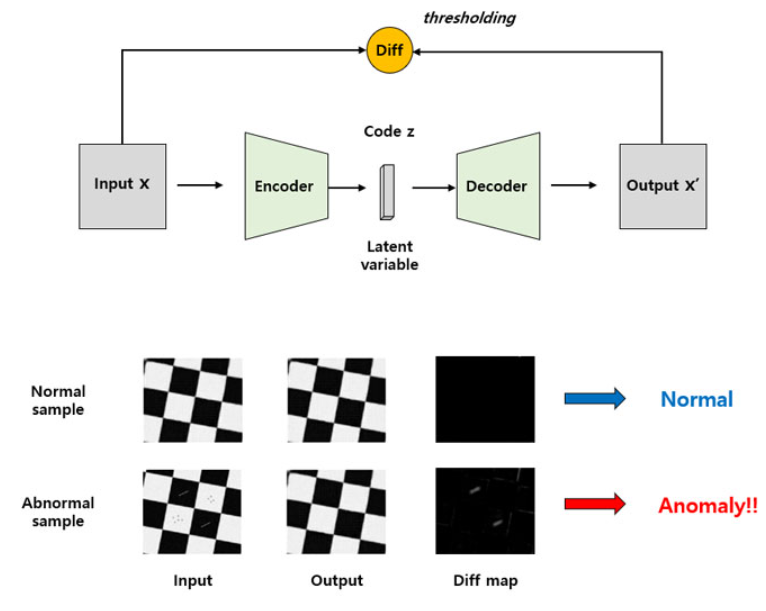
**Abstract**

이상치 탐지 모델 개발의 목적은 궁극적으로 FPR(False Positive Rate)을 최소화하고 보다 정확한 성능을 발휘하는 모델 및 시스템을 구축하는 것이다. 본 연구에서는 다변량 Time-Series에서 변화점 탐지를 수행하여 이상치 탐지 성능을 높이는, 규칙 기반 결정 시스템을 제안하고자 하며, 일관적이고 보편적인 모델 구축을 위해 혼재 구간에서의 결정 방법을 보다 디테일하게 제안한다.

현재 많은 이상치 탐지 모델 알고리즘은 변화점을 이상치로 분류하는 경향이 있기 때문에 높은 거짓 양성율을 갖는다. 이는 변화점의 데이터 포인트가 주위 데이터와 다른 분포를 뛰어 일어나는 현상으로 모델 정확도 향상의 주된 저해요소가 된다. 따라서 이를 해결하기 위해 이상치와 변화점을 매칭하는 과정을 거침으로써 거짓 경보(False alarm)를 줄일 수 있게 된다.

**Idea : Anomaly + Time Series Continuity + my Idea = High Accuracy & Low FPR**

**Anomaly Detection Algorithm : AutoEncoder(AE)**



오토인코더(이후 AE)는 어떠한 지도 없이도(즉, 레이블되어 있지 않은 훈련 데이터를 사용해서) 잠재표현(latent representation) 또는 코딩(Coding)이라 부르는 입력 데이터의 밀집 표현을 학습할 수 있는 인공 신경망이다. AE가 대부분의 데이터가 정상 Sample이라는 가정 하에 데이터의 주요 특성을 학습하여 정상 데이터와 비정상 데이터를 구분할 수 있다면, 비지도 이상치 탐지 알고리즘의 역할을 수행할 수 있다.

AE는 입력을 받으면 효율적인 내부 표현으로 바꾸고 입력과 가장 가까운 복원 결과를 출력한다. AE의 신경망 구조는 크게 두 부분으로 이루어지는데, 입력을 내부 표현으로 바꾸는 인코더(Encoder)와 내부 표현을 출력으로 바꾸는 디코더(Decoder)로 구성된다. 모델의 단점으로는 압축 정도(= code size = latent variable의 dimension) 같은 하이퍼 파라미터에 따라 전반적인 복원 성능이 좌우되기 때문에 다소 불안정하다는 단점이 있다. 또한, AE에 넣어주는 Input과 Output의 차이를 어떻게 정의할 것인지, 어떤 loss function을 사용할 것인지 등 여러 요인에 따라 성능이 크게 달라질 수 있음으로, 성능에 영향을 주는 요인들이 다수 존재한다.

AE는 위에서 언급했듯 대표적인 비지도 이상치 탐지 알고리즘이다. 학습된 AE에 정상 Sample을 넣어주면 Input과 Output에서 차이가 거의 발생하지 않는 반면, 비정상적인 Sample을 넣으면 AE는 정상 Sample처럼 복원하기 때문에 Input과 Output의 차이를 구하는 과정에서 차이가 두드러지게 발생한다. 여기서 Input과 Input을 복원한 Output의 차이를 재구성 오차(Reconstruction Error)라고 하며 주로 MSE, Cross-Entropy이다. 이 재구성 오차를 가지고 설정한 임계치에 따라 임계치를 넘는 데이터를 이상치로 판별한다.

따라서 AE를 활용한다면, 비지도학습 기반의 이상치 탐지 시스템 구축이 가능하다. 딥러닝 기반의 이상치 모델 중 AE기반 모델의 대표적인 예는 다음과 같다. (AE, VAE, Denoising AE, Spares AE)

**- Reference**

테이블이(가) 표시된 사진

자동 생성된 설명