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

**Paper**

**1. MDPI symmetry (2.713) : 심층 재구성 오류를 이용한 다중 도메인 시계열에서의 비지도 이상치 탐지 접근법**

텍스트이(가) 표시된 사진

자동 생성된 설명

1. **Abstract**

시계열 데이터에 대한 자동 이상 탐지는 사기 감지, 장애 진단 및 환자 모니터링과 같은 다양한 현실 영역에서 매우 중요하다. 현재의 이상치 탐지 기법들은 실제 이상 징후 중 현저하게 낮은 비율을 정확하게 감지한다. 게다가 대부분의 데이터셋은 레이블을 제공하지 않으며, 지도 학습 기반의 접근을 요구한다. 이러한 문제점들에 주목하여, 우리는 일괄 및 실시간 이상치 탐지에 적용할 수 있는 시계열 데이터에 대한 새로운 딥러닝 기반 비지도 이상치 탐지 접근법(RE-ADTS)을 제안한다. RE-ADTS는 시계열 재구성기와 이상치 탐지기를 포함하는 두 개의 모듈로 구성된다. 시계열 재구성기 모듈은 자기회귀(AR) 모델을 사용하여 최적의 창 너비를 결정하고 폭에 따른 후속 분석을 준비한다. 그런 다음 심층 오토인코더(DAE) 모델을 사용하여 데이터 분포를 학습시킨 다음, 시계열 데이터를 정상 데이터로 재구성하는데 사용한다. 이상 데이터일 경우 그들의 재구성 오류(RE)가 정상 데이터보다 높게 측정된다. 이 모듈의 결과로, RE와 차후의 압축된 표현형이 추정된다. 이후에 이상치 탐지 모듈은 RE 기반의 임계치를 사용하여 해당 시계열을 정상 혹은 이상치로 정의한다. 일괄 이상치 탐지를 수행할 경우, 밀도 기반 클러스터링 기법과 이상 임계치의 조합이 사용된다. 실시간 이상치 탐지의 경우, 클러스터링 과정 없이 오로지 이상 임계치만 사용된다. 우리는 일괄 및 실시간 이상치 탐지를 위해, 공개적으로 사용 가능한 총 52개의 시계열 벤치마크 데이터셋에 대해 총 두 가지 유형의 실험을 수행하였다.

* 시계열 데이터에 대한 딥러닝 기반 비지도 이상치 탐지 접근법(RE-ADTS)을 제시.

RE-ADTS는 시계열 재구성기와 이상치 탐지기 두 개의 모듈로 구성됨.

시계열 재구성기 모듈은 자기회귀(AR) 모델을 사용하여 최적의 창 너비를 결정하고 폭에 따른 후속 분석을 준비.

그런 다음 심층 오토인코더(DAE) 모델을 사용하여 데이터 분포를 학습시킨 다음, 시계열 데이터를 정상 데이터로 재구성.  
이상 데이터일 경우 그들의 재구성 오류(RE)가 정상 데이터보다 높게 측정됨.

1. **Introduction**

컴퓨터 하드웨어와 소프트웨어의 빠른 발전으로 인해, 시계열 데이터는 매분 매초마다 발생되고 있다. 시계열은 시간 순서로 저장된 데이터 포인트의 시퀀스를 의미하며, 각 포인트는 특정 기간 동안 정해진 시간 간격으로 측정된다. 이러한 데이터로부터 효과적인 분석을 통해 숨겨진 희귀 정보를 추출하는 것이 가능하다.

시계열에 대한 이상치 탐지는 많은 산업에 걸쳐 중요한 일이다. 대부분의 데이터셋과 호환되지 않는 비정상적인 패턴을 ‘novelty’, ‘outlier’, ‘anomaly’ 라고 부른다. 이상치는 다양한 클래스, 자연 변동, 데이터 측정 혹은 수집 오류와 같은 여러 가지 원인으로 인해 발생할 수 있다. 일부 실수를 통해 이상치가 발생할 수도 있지만, 때로는 극단적 날씨나 상당한 수의 트랜잭션을 나타낸다. 최근 몇 년 동안, 머신러닝 기반의 접근법이 많은 도메인 상에서 비정상 패턴을 감지하는데 널리 이용되었다. 예를 들어, 네트워크 보안 시스템의 악성 소프트웨어 및 침입 탐지, 제조의 결함 감지, 보안의 사기 탐지, 의료 진단의 질병 감지 등이 있다.

일반적으로, 머신러닝 기술들은 세 개의 카테고리로 나뉜다 : 지도, 준지도 그리고 비지도 학습. 시계열의 이상치 탐지는 비지도 학습 작업이다. 현재 비지도 이상치 탐지 접근법은 적은 수의 실제 이상 징후와 많은 수의 잘못된 정상 징후를 탐지하는 경향이 있다.

본 연구에서는 심층 오토인코더의 재구성 오류를 기반으로 하는 비지도 이상치 탐지 접근법인 RE-ADTS(Reconstruction error based anomaly detection in time-series)를 제안한다. AE는 출력값을 입력값과 최대한 동일하게 표현하는 방법을 학습하는 신경망의 한 종류이다. 일반적으로 데이터 차원을 줄이거나 노이즈를 소거하는데 사용된다. 그러나, 우리는 두 가지 목적으로인해 심층 AE모델의 RE 측정결과를 추가로 사용하였다 : 이상치를 탐지하고, 새로운 특성으로 사용함.

RE-ADTS는 아래와 같이 요약 가능하다.

먼저 AR 모델을 사용하여 선택한 최적의 창 너비를 기준으로 서브 시퀀스를 준비한다. 그 후, 준비된 서브 시퀀스는 심층 AE 모델에 의해 2차원 공간으로 압축된다. 이상치 탐지 모듈에서 첫째로 RE 기반의 이상 임계치를 압축 데이터셋의 재구성 오류에서 계산한다. 일괄 이상치 탐지에 경우, 압축된 서브 시퀀스는 DBSCAN 알고리즘에 의해 그룹화되고 클러스터 안에 각 서브 시퀀스의 RE가 이상 임계치와 비교된다. 만약 대부분의 서브 시퀀스가 이상 임계치를 초과하면, 그 클러스터는 이상치로 간주된다. 실시간 이상치 탐지의 경우에는, 클러스터링 과정 없이 서브 시퀀스의 RE가 이상 임계치와 비교된다.

* 본 연구의 기여는 아래와 같다.

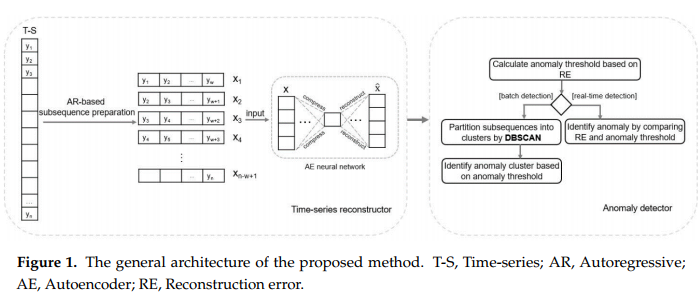
1. 비지도 상태에서 시계열 데이터의 이상치를 탐지할 수 있는 RE-ADTS 방법 제안.
2. RE-ADTS는 일괄 및 실시간 이상치 탐지 모두 이용 가능하다.
3. 자동으로 창의 크기를 결정함으로써, RE-ADTS는 수동 조작없이 다양한 도메인에 바로 적용될 수 있다.
4. 52개의 대표적인 시계열 데이터셋에서 공개적으로 사용 가능한 10개의 최첨단 이상치 탐지 방법을 평가했다. RE-ADTS는 대부분의 경우에서 다른 방법보다 더 높은 성능을 제공했다.
5. **Related Work**

시계열 데이터에서의 이상치 탐지는 데이터 사이언스 분야에서 가장 도전적인 문제 중 하나이다. 위에서 언급했듯, 우리는 이상치 탐지를 수행하기 위해 데이터 라벨 관점에서 세 개의 서로 다른 기술을 사용할 수 있다. 우리는 그 중에서도 비지도 이상치 탐지 접근법을 소개하고자 한다. Jiang et al은 클러스터링 기반의 비지도 침입 탐지 알고리즘(CBUID)을 제안했다. 이 알고리즘은 데이터셋을 클러스터 별로 그룹화하고, ‘클러스터에 속한 데이터 / 전체 데이터’의 비율을 기준으로 ‘normal’ or ‘attack’으로 클러스터를 라벨링하는 식으로 작동한다. Rousseeuw et al는 데이터 포인트에서 PCA 부분 공간까지의 직교 거리와 마할라노비스 거리를 기반으로 한 거리 점수를 사용하여 PCA 기반의 이상치 탐지 접근법을 소개했다. 정상 데이터에 대해선, 직교와 이상치 점수 거리 모두 작게 나타났다. 그리고 이상치 탐지를 위해 제안된 OC-SVM 기반의 준지도, 비지도 접근법이 있다. 먼저, time-delay 내장 프로세스를 사용하여 시계열 데이터를 벡터 집합으로 변환한 다음 OC-SVM을 이 벡터에 적용하면 된다.

최근에는 다차원 시계열 데이터에서 사용 가능한 딥러닝 기반의 이상치 탐지 방법이 제안되었다. 먼저, 그들은 2D Convolutional AE와 long short term memory AE를 기반으로 제안된 두 가지 버전을 제시했다. 그들은 이상치 탐지를 시작하기 전에 통계적 관점으로 특성들을 추가했다. 증축된 시계열과 이 증축 데이터의 재구성된 변형 간의 차이가 바로 이상치 지표로 사용된다. 그들은 재구성 오류 기반의 상위 5%를 이상치로 간주했다.

1. **Proposed Method**

오늘날 심층 신경망의 사용은 객체 인식, 감정 인식, 금융 이상 감지, 금융 시계열 예측 등 많은 영역에서 급격히 증가하고 있다. 우리는 시계열 데이터에 대해 새로운 비지도 이상치 탐지 접근 방식을 확장하여 딥러닝 기반의 이상치 탐지 방법을 제안한다. 지난 접근법에서 심층 AE 모델의 입력은 이전 값에 의존되지 않는다. 더욱이, 심층 AE 모델은 전체 데이터 집합이 아닌 거의 정상에 가까운 하위 집합에 대해 훈련되었다. 따라서 데이터셋이 시간 간격 없이 정렬되어 있어야 하기 때문에 시계열 분석에서 이 방법을 직접 사용할 수는 없다. 본 연구에서 제안된 접근방식은 이러한 문제를 다루기 위해 설계되었으며, 시계열 재구성기와 이상치 탐지 모듈로 구성된다.



시계열 재구성기 모듈에서, 우리는 심층 AE와 AR 모델을 결합하였다. AE는 입력을 저차원 공간에 투영한 다음 다시 출력으로 재구성하는 신경망의 한 유형으로 이 특성을 사용하여 RE 기반의 이상치 평가를 수행할 수 있다. 다시 말해, AE 모델이 모든 서브 시퀀스로부터 학습을 할 때 이상치 보다 정상에 대해 더 많이 나타난다. 그 이유는 전체 데이터에서 이상치의 비율이 매우 작기 때문이다. 이상치 탐지 모듈에서는 일괄 이상치 탐지를 위해 RE 기반의 이상 임계치를 가진 밀도 기반 클러스터링 기술이 사용된다. 실시간 이상치 탐지의 경우에는 클러스터링 기술 없이 RE 기반의 이상 임계치만 사용된다.

예를 들어, 우리가 1,000개의 시간 정렬 관측치를 가지고 있다고 가정하자. AR 모델에서 도출된 최적의 서브 시퀀스 길이는 15이고, 총 986개의 서브 시퀀스가 준비된다. 우리는 심층 AE모델을 사용하여 15차원 공간으로부터 이러한 시퀀스를 잠재 변수와 RE로 구성된 2차원공간으로 변환한다. 그러고 나서, 모든 서브 시퀀스의 RE를 통해 이상 임계치를 측정한다. 일괄 이상치 탐지에 경우, 986개의 서브 시퀀스의 2차원 표현은 밀도에 기초하여 그룹화된다. 결과적으로 총 3개의 그룹이 생성되었고 각각 100, 500, 350의 서브 시퀀스가 할당되었으며 36개의 서브 시퀀스가 그룹화되자 않았다고 가정한다. 우리는 클러스터링 되지 않은 36개의 서브 시퀀스를 하나의 클러스터로 간주하고, 이상 유무에 관계없이 총 4개의 클러스터를 확인할 것이다. 1번 클러스터는 100개의 서브 시퀀스로 구성되어 있지만, 그 중 89개의 RE가 이상 임계치보다 높았다. 다른 클러스터의 경우에는 5~15%의 서브 시퀀스만 임계치를 초과했다. 따라서, 1번 클러스터만 이상치로 명명하였다. 실시간 이상치 탐지의 경우 각 서브 시퀀스의 RE를 클러스터링 프로세스 없이 하나씩 이상 임계치와 비교했으며, 임계치보다 RE가 높은 모든 서브 시퀀스를 이상치로 표시하였다. 예를 들어, 위에서 언급한 1번 클러스터의 89개의 서브 시퀀스는 이상치로 여겨진다.

* 1. **Autoregressive Based Deep Autoencoder Model for Time-Series Reconstruction**

아래 본문에서는, 도메인 지식없이 시계열 데이터로부터 자동적으로 이상치 탐지를 할 수 있는 방법에 대해 다루도록 하겠다. 그러므로, 우리가 제안한 모델의 파라미터 조정은 완전이 자동화되었다. 사실 시계열 분석의 성과는 어떻게 서브 시퀀스를 조정하는지에 달려있다. 달리 말해, 최적의 길이의 서브 시퀀스를 선택하는 것이 시계열 분석에서 높은 성과를 얻는데 매우 주요하다. 그러나, 시계열 데이터를 위한 일반적이고 고정된 창의 너비는 정해져 있지 않으며, 각 시계열 데이터는 개별적인 특성을 가지고 있다. 창 너비를 선택하는 것은 쉬운 일이 아니다. 만약 창이 너무 크면, 분석을 위해 더 많은 과거 데이터를 요구하기 때문에 계산 복잡도와 시간 지연을 증가시킨다. 반대로 너무 작으면, 시계열의 패턴을 놓치게 된다. 우리가 제안한 방법에서는, 특정 타임스탬프의 값과 이전 값으로 구성된 서브 시퀀스의 재구성 오류가 이상치 탐지에 사용되었다. 다시 말해, 특정 시점이 이상치 인지 아닌지 결정하기 위해 우리는 서브 시퀀스 내에서 그 이전의 값들을 사용하였다. 그러나 사용자가 최적의 주기성을 지정해야 하기 때문에 모든 경우를 하나씩 시도해 보는 것은 비용이 많이 들게 됩니다.

대부분의 이상치 탐지 접근법에선, 사용자가 지정한 창 길이가 사용된다. 따라서 이러한 메커니즘은 변화하는 데이터셋에 적합하지 않다. 창 너비의 선택은 AR 모델의 차수를 결정하는 접근법으로 해결이 가능하다. AR 모델은 시계열 분석을 위한 가장 성공적이고 유연하며 사용하기 쉬운 모델이다. RE-ADTS는 최적의 창 너비를 사용한 슬라이딩 윈도우 방법으로 전체 시계열 데이터의 서브 시퀀스를 준비한다. 서브 시퀀스의 길이는 AR 모델을 통해 결정되며, AR 모델은 이전 값에 선형적으로 의존하는 출력 값을 출력한다. 즉, ‘t’라는 시간의 값은 이전 ‘p’ 기간의 값에 따라 달라지며 여기서 ‘p’를 차루라고 한다.

차수가 p인 AR 모델의 수식은 아래와 같다.



위 식에서 e는 노이즈, c는 상수, q1~qp는 모델 파라미터 그리고 yt-1~yt-p는 이전 데이터의 값을 의미한다. AR 모델의 과정에는 불확실성의 정도가 있으며, 최적의 창 너비(차수P)를 결정하기 위해 모델 선택법이 사용되었다. 가장 적절한 창 너비를 선택하기 위해 일반적으로 사용되는 접근법은 AR (p) 모델을 서로 다른 수의 차수로 설정하고, 모델 선택 기준의 무언가를 최소화하는 p 값을 선택하는 것이다. 우리는 차수를 선택하기 위해 CV, AIC, BIC 그리고 T-STAT와 같은 몇 가지 기준을 사용하였다.

AIC는 Akaike가 제시한 모델 선택법이다. 여기서 p는 차수이며, e^p는 2차 손실을 기준으로 한 평균 예측 오차, n은 표본의 개수이다



BIC는 또 다른 모델 선택 기준으로, AIC와의 유일한 차이점은 AIC 페널티의 상수 2개가 데이터셋 크기의 로그형태로 대체된다는 것이다. 여기선 아래 식을 최소화하는 모델을 선택한다. 여기서 p는 차수이며, e^p는 2차 손실을 기준으로 한 평균 예측 오차, n은 표본의 개수이다.



위에서 제시된 접근 방식은 자동으로 창 너비를 선택하는 과정이고, 이를 통해 다양한 시계열의 특성에 적응할 수 있다. 아래 있는 표 A1은 고정된 최적의 창 너비를 사용하여 각 도메인에 대한 평균 f-검정의 결과를 비교한 것이다.

테이블이(가) 표시된 사진

자동 생성된 설명

준비된 서브 시퀀스는 재구성 오류를 구하기 위해 심층 AE 모델을 통해 재구성 된다. AE는 입력 값을 복사하는 방법을 학습하여 출력 값과의 차이를 가능한 줄이는 비지도 학습 알고리즘이다. AE의 구조는 인코더와 디코더로 이루어져 있으며, 두 구성 모두 뉴런으로 된 신경층을 가지고 있다. 각 뉴런은 이전 레이어에서의 가중치 합 결과를 전환한 활성화 함수 출력 값을 받게 된다. AE는 입력과 재구성된 결과의 차이를 줄이기 위해 각 뉴런의 가중치 값을 바꿔가며 학습한다. 인코더 부분은 vector x를 입력으로 받고, 잠재 공간이라 부르는 압축 표현 h로 벡터를 투영한다. 그러고 난 후, 잠재 공간은 디코더 부분에서 재구성된 vector x로 다시 재구성된다. 아래 식은 AE의 과정을 묘사한 것이다.

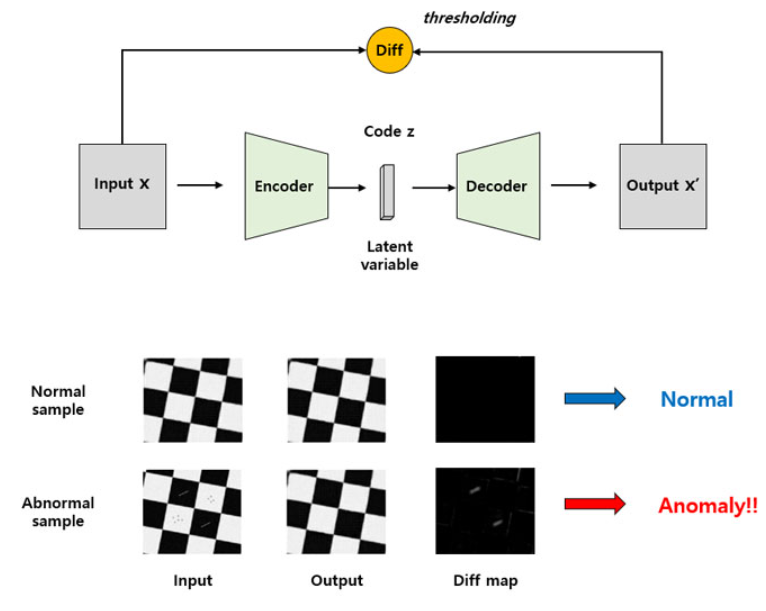


1. **Experimental Study**

우리는 52개의 대표 시계열 데이터를 포함한 6개의 서로 다른 도메인에서 RE-ADTS를 평가했다. 그리고 그 결과를 10개의 다른 이상치 탐지 방법과 비교하였다.

**0. Abstract**

**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