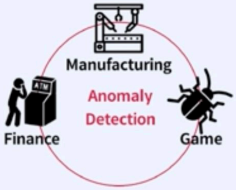
이상 탐지강의 기획 의도 및 목적

이상 탐지 강의 학습 목표 및 수강 후 기대효과

· 기업 Needs



① 제조

- 설비 예지 보전으로 고장 발생 시 Loss 비용 절감

② 금융

- FDS(Fraud Detection System)을 통해 이상거래를 탐지하고 사기 거래로 발생할 수 있는 리스크 절감

③ 게임

- 재화, 몬스터, 아이템 등 게임 내 발생할 수 있는 이상 현상 방지를 통해 어뷰징 해결

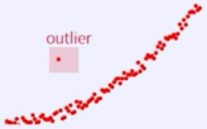
이상 탐지 정의 및 분석가치

기업에서 이상 탐지를 하는 이유 및 분석적 가치

· 이상치(outlier) vs 이상(abnormal)

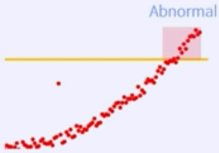
① 이상치(outlier)

- 관측된 데이터의 범위에서 많이 벗어난 아주 작은 값이나, 큰 값
- 분석하고자 하는 데이터에서 적은 확률로 나타나는 데이터
- 분석 결과 해석 시 오해를 발생시킬 수 있기 때문에 사전 제거



② 이상(abnormal)

- 문제해결의 관점
- 현업의 Domain 관점에서 보았을 때, 문제발생 가능성이 높은 데이터
- 정상적인 범주에 데이터라도 이상으로 정의할 수 있음
- 일반적으로 자주 발생하지 않는 패턴이 이상일 확률이 높음



이상 탐지 정의 및 분석가치

기업에서 이상 탐지를 하는 이유 및 분석적 가치

- 이상치(outlier)란 데이터 관점, 이상(abnormal)이란 현업의 문제해결 관점 View
- 이상 탐지(Anomaly Detection)는 이상이라고 정의한 사건 및 패턴을 탐지하는 활동
- 기업에서 이상 탐지를 하는 목적

① [제조] 심각한 고장 발생 전 이상 탐지를 통해 심각한 Risk 방지
② [금융] 비정상적 거래 및 사기 거래 방지를 통해 소비자 보호
③ [게임] 버그 유저 및 비정상 유저 탐지를 통해 게임 정상 운영

「더 큰 Risk가 발생하기 전 피해를 최소화 하기 위함」

이상 데이터 발생원인

다양한 Case의 이상 데이터 발생원인

- 자연적으로 발생하는 이상(abnormal) 사건 이외에도 다양한 원인에 의해 이상 발생
- 이상치(Outlier) 데이터를 분석하기 사전 점검을 통해 발생 원인 제거 가능
- 본격적인 데이터 분석 적 올바르게 데이터가 수집되어 있는 상황인지 사전 점검 필수

① 표본추출 오류 (sampling error)

- 데이터를 샘플링 하는 과정에서 잘 못 샘플링 한 경우 ex) 다른 모수에서 추출하는 경우

② 입력 오류 (data entry error)

- 데이터를 수집, 기록, 입력하는 과정에서 발생하는 Human error, 데이터 분포로 쉽게 탐지 가능

③ 실험 오류 (experimental error)

- 실험을 통해 데이터를 수집하는 경우, 실험 조건이 동일하지 않은 경우 ex) 챔버A != 챔버B

이상 데이터 발생원인

다양한 Case에 이상 데이터 발생원인

④ 측정 오류 (measurement error)

- 측정 장비를 통해 데이터를 수집하는 경우, 측정기 자체에 오류가 발생한 경우

  ex) 다른 동일 측정기와의 측정 분포를 확인함으로써 오류를 찾아낼 수 있음

⑤ 데이터 처리 오류 (data preprocessing error)

- 데이터 ETL(Extract > Transform > Load) 과정에서 발생할 수 있는 데이터 처리 오류

⑥ 자연 오류 (natural Outlier)

- 자연스럽게 발생하는 이상 값 (※ 우리가 데이터 분석을 통해 해결하려고 하는 문제)

이상 탐지의 종류 (1) - Type

데이터 유형에 따른 이상 탐지의 종류

· Data Type

① Time series(sequential) vs static(정적인, Point)

② univariate(단변량) vs multivariate(다변량)

Time	Col1

vs

Time	Col1	Col2	Col3

③ data type (binary / categorical / continuous / hybrid)

1,0 A, B, C 10.6

④ relational(상관관계가 있는) vs independent(독립적인)

⑤ well-known or not (기존 룰의 적용 가능한 / 알려져 있지 않은)

이상 탐지의 종류 (1) - Type

데이터 유형에 따른 이상 탐지의 종류

① Point Anomaly Detection

- 축적된 시간 동안 정적인 점 분포에 초점
- 특정 Point의 이상치를 감지
- 일반적으로 말하는 데이터 내 outlier

② Contextual Anomaly Detection

- 시계열과 같은 동적인 특성에 초점 (※ 과거에 데이터가 현재의 값에 영향을 미칠 때, sequential)
- 연속적인 변화 패턴을 읽어 이상치 감지, 맥락을 고려해서 예상 변화와 동떨어진 결과를 탐지
- 민감하면 정상상황에서도 이상탐지가 되고, 둔감하게 만들면 비정상 상황에서 이상탐지를 놓칠 수 있음

③ Collective Anomaly Detection

- Contextual Anomaly와 다르게 Global 상황에서 변칙적인 이상치가 아닌, Local한 이상치
- 개별 인스턴트가 아닌 집합 인스턴트 비교를 통해 이상을 확인

④ Online Anomaly Detection

- 실시간 데이터 수집 체계가 구축되어 있는 환경에서 탐지
- 실시간 데이터를 어떻게 빠르게 처리하고 이상을 탐지할지 설계하는 것이 매우 중요

⑤ Distributed Anomaly Detection

- 관측치의 정상 분포에서 벗어나는 이상 데이터를 탐지

이상 탐지의 종류 (2) - Label

데이터 Label(정답) 유무에 따른 이상 탐지 방법론

· 현재 보유하고 있는 데이터의 성격에 따라 이상 탐지의 종류 및 해결 방법론이 결정

① Supervised Anomaly Detection

- 모든 관측치에 대해서 이상(abnormal)과 정상(normal) 라벨이 붙어있는 경우, 자주 발생하지 않는 이벤트인 이상(abnormal)에 대해서 정의할 수 있는 상황이며, 충분히 학습할 만한 이상(abnormal) 데이터가 확보되어 있는 상황
- 일반적인 이진분류(Binary Classification) 문제로 해결할 수 있음
  (※ 이상(abnormal)의 종류가 다양하다면 Multi-Classification 문제로 해결)
- 이상(abnormal) 데이터 확보가 가능하나 적은 수의 데이터가 존재할 확률이 있으므로, 소수 클래스 문제, 불균형 문제를 해결해야 하는 상황을 고려할 수 있음
- 성능평가 시 ACC(Accuracy) 관점이 아닌, Recall 관점에서의 성능평가도 고려해야 함
  이상(abnormal) 데이터가 매우 적기 때문에 ACC는 높게 산출될 가능성이 큼

· 장점 : 양/불 판정 정확도가 높다
· 단점 : Class-Imbalance 문제

이상 탐지의 종류 (2) - Label

데이터 Label(정답) 유무에 따른 이상 탐지 방법론

② Semi-Supervised Anomaly Detection

- 이상(abnormal)에 대한 라벨이 없으며, 정상(normal) 데이터만 보유하고 있는 상황, 충분히 학습할 만한 이상(abnormal) 데이터가 없을 시 정상(normal) 데이터만을 사용하여 이상(abnormal)을 탐지
- 정상(normal) 데이터의 패턴을 학습하여 이상(abnormal) 데이터가 들어올 시 정상(normal) 데이터와 차이점을 파악하여, 이상(abnormal) 데이터를 탐지
- 정상 데이터를 둘러싸고 있는 discriminative boundary(구분선)를 설정하고, 이 boundary를 최대한 좁혀 밖에 있는 데이터들을 이상(abnormal)로 간주하는 것
- 대표적인 방법으로는 One-Class SVM이 존재

- **장점** : 연구가 활발하게 이뤄지고 있고, 정상(normal) 데이터만 있어도 학습이 가능
- **단점** : Supervised Anomaly Detection 방법론 대비 양/불 판정 정확도가 떨어짐

이상 탐지의 종류 (2) - Label

데이터 Label(정답) 유무에 따른 이상 탐지 방법론

③ Unsupervised Anomaly Detection

- 정상(normal)과 이상(abnormal)에 대한 라벨이 모두 존재하지 않는 경우
- 모든 데이터를 활용하여 학습
- 일반적으로 딥러닝 알고리즘을 사용하여 정상 패턴을 학습하고, 학습된 모델을 통해 이상(abnormal) 데이터를 탐지함

- **장점** : 라벨링 과정이 필요 없음
- **단점** : 양/불 판정 정확도가 높지 않고 hyper parameter에 매우 민감함

이상 탐지의 종류 (3) - Train Data

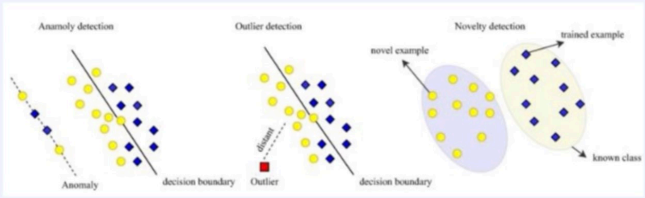
해결하려는 문제의 목적과 학습 데이터에 따른 이상탐지의 종류

· 새로운 관측치가 기존 분포In-distribution에 속하는지, 기존 분포를 벗어 났는지Out-of-distribution를 구분

· 학습할 데이터를 어떻게 정의하는지에 따라 문제의 성격과 해결 방법론이 달라짐

① Outlier Detection

- 학습 데이터를 통해 정상 데이터의 범위를 결정하고, 이를 초과할 시 이상(abnormal) 데이터로 간주
- 훈련 데이터 셋에 정상 샘플과 이상치 샘플을 모두 포함하고 있다. (대부분 정상 샘플)
- Training data가 Outlier를 포함하고 있는 상황에서 Training data에서 Central mode에 Fit하는 것
  (중앙에서 벗어난 데이터를 Outlier로 판단.)



이상 탐지의 종류 (3) - Train Data

해결하려는 문제의 목적과 학습 데이터에 따른 이상탐지의 종류

② Novelty Detection

- 학습할 데이터가 이상치(Outlier)에 의해 오염되지 않았다고 가정
- 우리는 새로운 패턴에 데이터를 찾는 것에 관심
- Out-of-distribution : 현재 보유하고 있는 In-distribution 데이터 셋을 이용하여 multi-class classification network를 학습시킨 뒤, test 단계에서 In-distribution test set은 정확하게 예측하고 Out-of-distribution 데이터 셋은 걸러내는 것을 목표
- 훈련 데이터 셋에 있는 모든 샘플과 달라 보이는 새로운 샘플을 탐지하는 것이 목적 (훈련 데이터에 포함된 샘플은 특이치로 생각하지 않는다.)
- Training data가 특이치 데이터를 포함되지 않은 채 학습하고 (사이킷런의 문서에서는 오염되지 않았다고 표현), 예측 시 새로운 관측치인지 확인한다. 따라서 알고리즘으로 감지하고 싶은 샘플들을 제거한 훈련 데이터 셋이 필요