Time-Series Aware Precision and Recall for Anomaly Detection

Won-Seok Hwang, Jeong-Han Yun, Jonguk Kim, Hyoung Chun Kim
The Affiliated Institute of ETRI

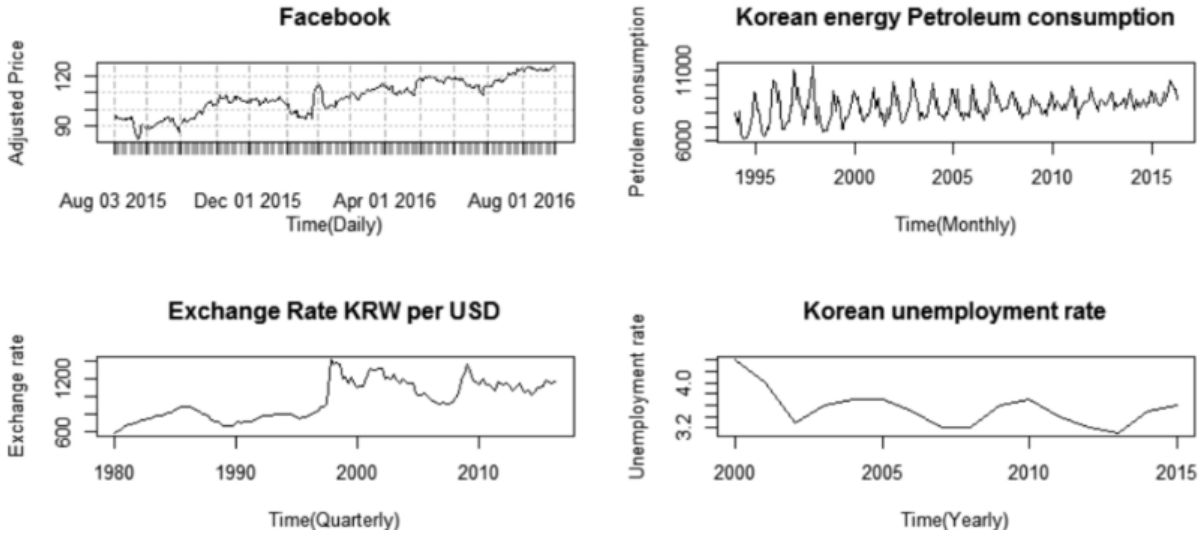
2024 07월 09일
박세준

Data Mining and Intelligence Systems
Chungang University

Background: Time Series Data

- **Time Series Data**
  - 일정한 시간 동안 수집 된 일련의 순차적으로 정해진 데이터 셋의 집합
  - 시간에 관한 순서가 존재
  - 연속한 관측치는 서로 상관관계 존재

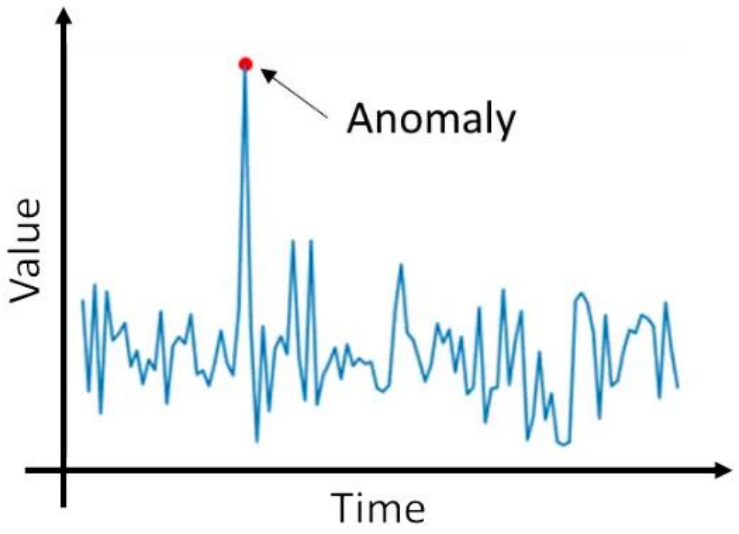
- **Example**

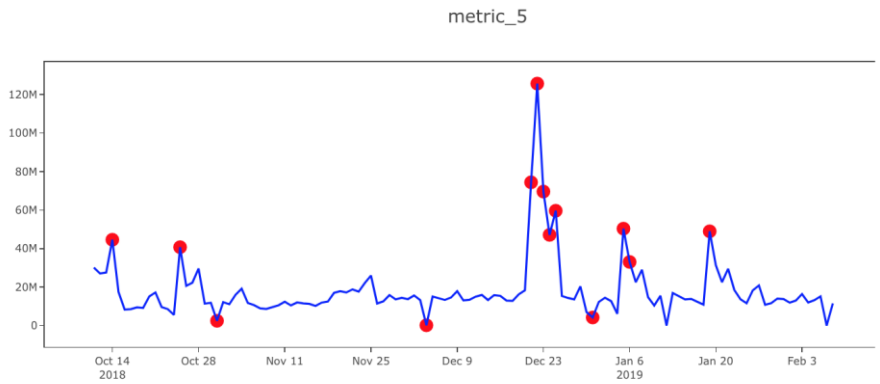


Background: Anomaly Detection

- **Anomaly Detection**
  - Data Set에서 정상적인 패턴과 일치하지 않는 비정상적이거나 예외적인 패턴을 식별하는 것

- **Example**





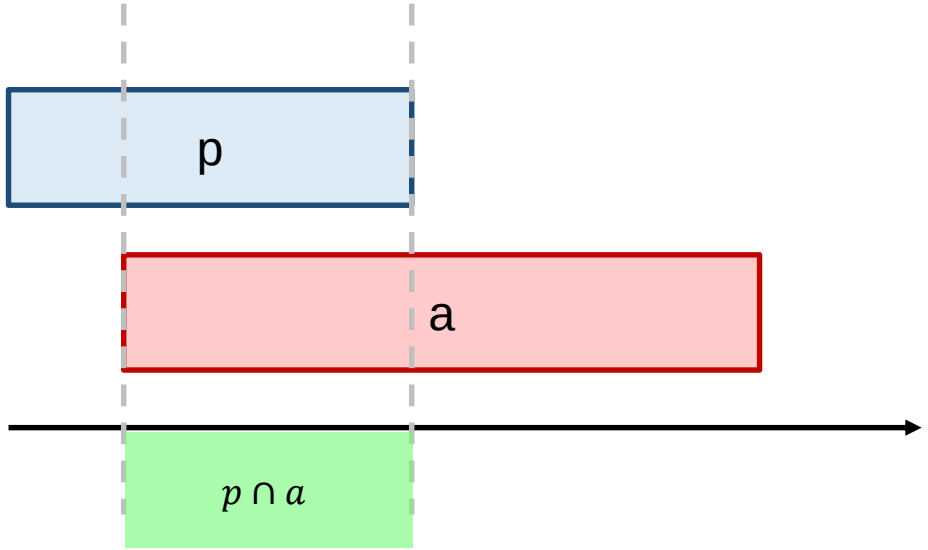
Background: Metrics

- **Instance based metrics**
  - 많은 연구에서 인스턴스의 anomaly detect와 prediction을 위해 사용한 방법으로 Precision, Recall, ROC curve, AUC등이 있음.
  - 긴 anomaly혹은 prediction에 영향을 받는 문제가 있음
  - 몇몇 연구에서 해당 metric은 time series data에 알맞지 않다는 주장이 있음

- **Time Series based metrics**
  - 정확한 prediction에 positive score를 주고 그렇지 않다면 negative score를 줌
  - Instance의 series가 아닌 하나의 instanc에만 적합하다는 문제가 있음
  - Scoring function과 magic number에서 애매한 부분들이 있음

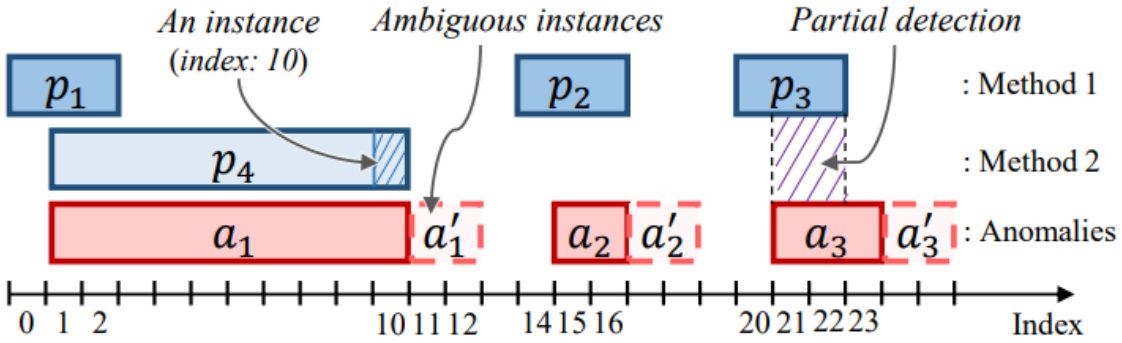
Background

- **Precision**
  - $\frac{|p \cap a|}{|p|}$ (맞힌 예측) / (전체 예측)
- **Recall**
  - $\frac{|a \cap p|}{|a|}$ (감지된 이상) / (전체 이상)



Background: Pre-Existing problems

- **First Problem**
  - Instance Based Metrics들은 Time Series Data에 적합하지 않음
    - 다양한 anomalies를 detect하는 것 보다 길이가 긴 anomaly를 detect 하는 것에 고점을 줌
- **Second Problem**
  - Ambiguous Instance들을 간과하였다
    - Anomalies들 만들기 위해 연구자들이 그들의 testbed를 비정상으로 조정하였을 때 그 남아있는 조정값들이 얼마나 오랬동안 영향을 줄지 추정하기 어렵다. 그렇기 때문에 anomaly뒤에 오는 것들은 label이 normal이어도 anomalous할 가능성이 존재한다.



Goal

- **First Goal**
  - 다양한 anomalies탐지에 높은 점수 부여
    - 기존의 metric들과 다르게 다양한 anomalies를 detect하였다면 높은 점수를 부여하고, 더 좋은 성능을 가진 것으로 간주할 것

- **Second Goal**
  - Ambiguous Instance 고려
    - 기존의 metric은 간과하고 넘어간 ambiguous instance들을 고려하도록 하여 정확도를 높일 것

First Goal

- **방법**
  - 간단하게 detect된 것들의 개수를 센다 하지만 부분적으로 감지한 것에 문제가 있음
  - 부분적으로 detect된 anomalies들을 평가하기 위해 detection score, portion score 두가지 score방법 사용

- **Detection Score**
  - Detect된 anomalies의 개수를 고려한다
  - 부분적으로 detect된 것들을 fully detect된 것으로 간주

- **Portion Score**
  - 각각의 anomaly에서 detect된 인스턴스의 비율이다

Second Goal

- **방법**
  - Subsequent scoring을 사용한다

- **Subsequent scoring**
  - 애매한 인스턴스들을 detect한 prediction에 점수를 할당
  - Anomaly와 거리가 멀 수록 점수는 떨어진다.
  - 언제나 anomaly보다 점수가 낮다

RR, RP

- **Range based precision(RP), recall(RR)**
  - 유일하게 같은 문제에 초점을 맞춘 metrics들이다.
  - 각각 4개의 aspects가 있다
    - Existence, size, position, cardinality
  - Existence는 만약 하나 혹은 하나이상의 instance가 detect되었다면 highest score를 할당한다
  - Size와 position은 각각 detect된 anomaly instances들의 숫자와 연관된 position을 의미한다
  - Cardinality는 하나 이상의 prediction과 연관된 anomaly를 penalize한다.
  - Existence로 인해 아주 작은 부분만 detect해도 극단적으로 높은 점수를 주는 문제가 있다
  - Position과 cardinality을 고려하는 적절한 함수를 결정하는 것이 어렵다

TaR, TaP

- TaR
  - Anomalies를 찾는 성능
- TaP
  - 오경보 발생 빈도
- **Novelty**
  - Detection과 Portion scores로 구성된 식
  - 연구자들이 목적에 기초하여 importance를 조절할 수 있다.
  - Ambiguous instances를 규명하기 위해 subsequent scoring을 고려하였다.

Time-Series Aware Recall(TaR)

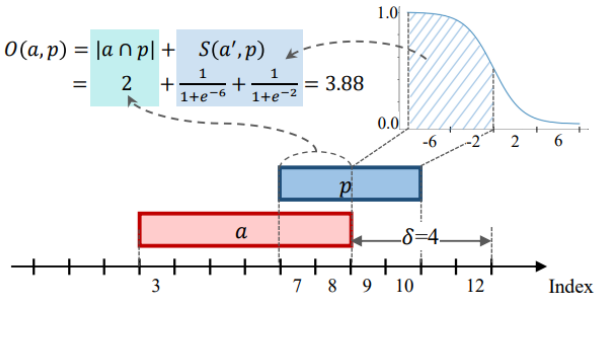
- **의미**
  - Anomalies를 찾는 성능

- **정의**
  - $TaR = \alpha \times TaR^d + (1 - \alpha) \times TaR^p$
  - α 는 비율을 조정하기 위해 사용
  - TaR^d는 전체 anomalies들의 집합에서 $A^d(\theta)$(detect된 anomalies)들의 집합

- TaR^d
  - $TaR^d = \frac{|A^d(\theta)|}{|A|}$ $where\ A^d(\theta) = \{a | a \in A\ and\ \frac{\sum_{p \in P} \mathbf{O}(a,p)}{|a|} \geq \theta\}.$
  - θ는 detection method가 anomlalies의 수가 많을 때 아주 적은 수의 인스턴스만 suspect했다면 anomaly를 detect하기 어렵기 때문에 threshold로 사용한다.
  - Overlap score $O(a,p)$는 예측 p 가 anomaly a를 찾을 확률을 의미한다.

Time-Series Aware Recall(TaR)

- $\boldsymbol{O(a,p)}$
  - $O(a,p) = |a \cap p| + S(a',p)$
  - p가 a의 인스턴스들 detect한 것에 정비례한다
  - p가 a뒤에 오는 ambiguous instance a'를 가리킨다면 a또한 찾을 수 있다
  - $|a \cap p|$: scores for detecting anomalous
  - $S(a',p)$: scores for detecting ambiguous instance
- $\boldsymbol{S(a',p)}$
  - ambiguous instance a'와 prediction p 를 위한 subsequent score
  - $\mathrm{S}(a',p) = \sum_{i \in (a' \cap p)} \frac{1}{1+e^{i'}} \; where \; i' = -6 + \frac{12(i - t_{a'} - 1)}{\delta - 1}$
  - $t_{a'}$는 ambiguous instance a'의 첫번째 인덱스이다
  - anomaly a 와 a'의 거리가 멀다면 anomalous할 가능성이 떨어진다.
  - δ는 ambiguous instance로 간주하는 instance의 개수
  - 위 논문에선 계산이 잘 되도록 inverse sigmoid함수를 사용하였고, 이 수식이 잘 된다면 어떤 함수를 사용하여도 상관없다



Time-Series Aware Recall(TaR)

- $\boldsymbol{TaR^P}$
  - 각 anomaly에 대하여 detect된 비율과 $O(a,p)$를 사용하여 subsequent score를 고려한다.
  - min함수는 최고점을 1로 제한하기 위해 1과 fraction중에서 작은 값을 도출한다.

$$TaR^p = \frac{1}{|A|} \times \sum_{a \in A} min\left(1, \frac{\sum_{p \in P} \mathbf{O}(a,p)}{|a|}\right).$$

Time-Series Aware Precision(TaP)

- **의미**
  - 오경보 발생 빈도
- $\boldsymbol{TaP}$
  - $TaP = \alpha \times TaP^d + (1-\alpha) = TaP^p$
  - TaR과 마찬가지로 TaP 도 TaP^d 와 TaP^p의 조합이다.
  - TaP^d는 전체 prediction P에서 $P^c(\theta)$(correct prediction)의 fraction이다.
- $\boldsymbol{TaP^d}$
  - $P^c(\theta)$는 correct instances가 차지하고 있는 비율이 θ보다 큰 predictions들로 구성되어있다.
  - $\boldsymbol{TaR^d}$와 같은 이유로 θ를 threshold로 사용한다.

$$TaP^d = \frac{|P^c(\theta)|}{|P|} \; where \; P^c(\theta) = \{p | p \in P \; and \; \frac{\sum_{a \in A} \mathbf{O}(a,p)}{|p|} \geq \theta\}$$

- $\boldsymbol{TaP^p}$
  - 각각의 prediction에 대하여 correct part의 평균적인 비율 $TaP^p = \frac{1}{|P|} \times \sum_{p \in P} \left(\frac{\sum_{a \in A} \mathbf{O}(a,p)}{|p|} \right)$

Evaluation: Result of Example

- **Fig1과 Fig3에서의 성능 비교**

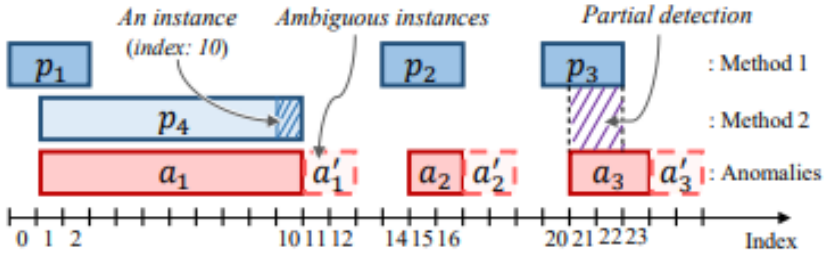


Figure 1: Example of inaccurate precision and recall in time-

Table 1: Results of Examples

Fig. / Method	Metric (Recall)			Metric (Precision)		
	Recall	*RR*	*TaR*	*Precision*	*RP*	*TaP*
1 / 1	0.40	0.87	0.64	0.67	0.84	0.84
1 / 2	0.67	0.33	0.33	1.00	1.00	1.00
3 / 3	1.00	1.00	1.00	0.38	0.56	0.19
3 / 4	0.00	0.51	0.01	0.03	0.54	0.23

  - TaR, RR에서 method 1이 method 2보다 더욱 좋은 점수를 받음
  - TaP, RP에서 method 1이 precision보다 더 높은 점수를 받은 이유는 instance가 아닌 predictions들의 점수의 평균을 점수로 하였기 때문이다.

Evaluation: Result of Example

- **Fig1과 Fig3에서의 성능 비교**

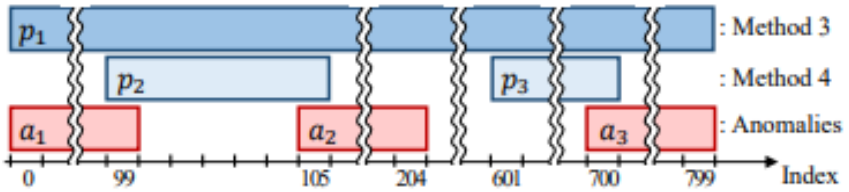


Figure 3: Examples of anomalies and predictions produced

Table 1: Results of Examples

Fig. / Method	Metric (Recall)			Metric (Precision)		
	Recall	*RR*	*TaR*	*Precision*	*RP*	*TaP*
1 / 1	0.40	0.87	0.64	0.67	0.84	0.84
1 / 2	0.67	0.33	0.33	1.00	1.00	1.00
3 / 3	1.00	1.00	1.00	0.38	0.56	0.19
3 / 4	0.00	0.51	0.01	0.03	0.54	0.23

  - TaP가 RP보다 Method 3과 4의 점수를 낮게 줌
  - TaR이 RR보다 Method 3과 4의 점수를 낮게 줌
  - TaP, TaR이 더 적합한 metric이다.

Evaluation: Result of SWaT dataset

- **SWaT Dataset에서의 성능**
  - iForest는 전체 228개중 66개만 탐지 했기에 TaR이 낮다.
  - Seq2Seq가 더 자주 ambiguous 한 instance들 까지 anomalies라고 detect했기 때문에 TaP가 RP보다 높다
    - Subsequent Score가 extra score를 줬기 때문

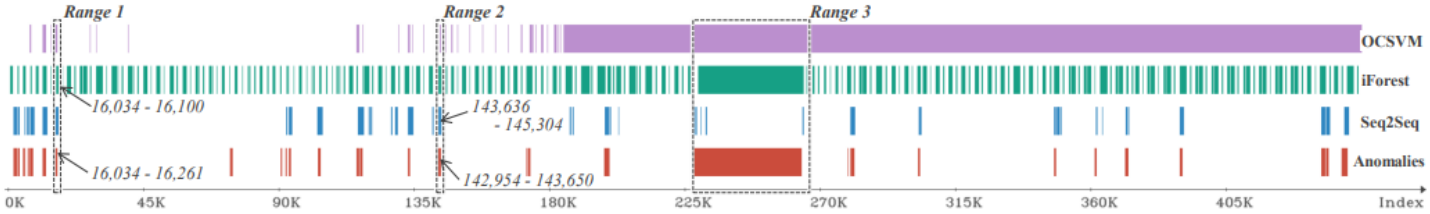


Table 2: Detection Results using SWaT dataset

Method	Metric (Recall)			Metric (Precision)		
	Recall	*RR*	*TaR*	*Precision*	*RP*	*TaP*
OCSVM	0.85	0.61	0.55	0.17	0.14	0.17
iForest	0.74	0.52	0.40	0.30	0.04	0.05
Seq2Seq	0.25	0.66	0.65	0.59	0.35	0.44

  - TaP, TaR이 더 적합한 metric이다

Conclusions

- **Portion Score**
  - 얼마나 정확하게 각각의 anomaly를 detect했는지
- **Detection Score**
  - 얼마나 많은 anomalies들을 detect했는지
- $\boldsymbol{TaP}, \boldsymbol{TaR}$
  - Portion score와 Detection score를 사용하여 만듬
  - Ambiguous한 instance들을 고려하기 위해 subsequent score를 사용
  - 기존에 존재하던 metric들 보다 훨씬 합리적인 방법이다.