Do You Know Existing Accuracy Metrics Overrate Time-Series Anomaly Detections?

Won-Seok Hwang, Jeong-Han Yun, Jonguk Kim, Byung Gil Min
The Affiliated Institute of ETRI Daejeon, South Korea

2024년 7월 9일

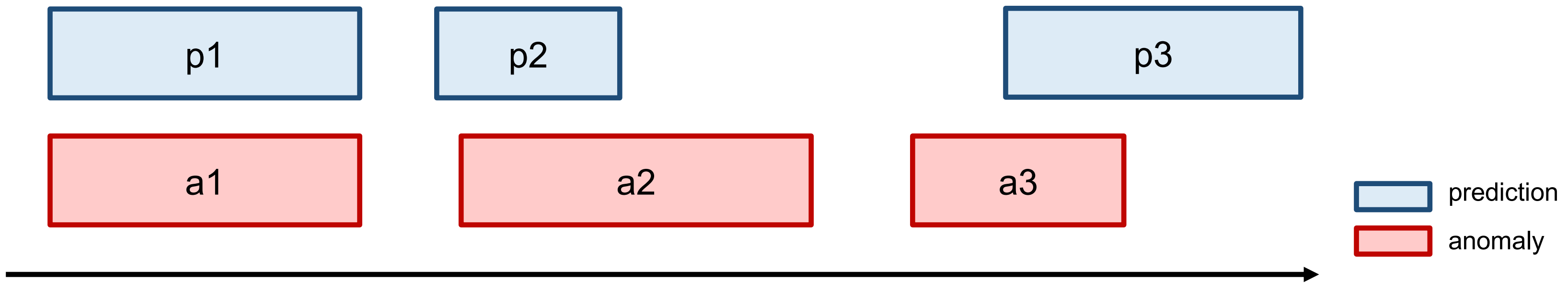
이규원

Department of Computer Science and Engineering
Chung-Ang University

Background

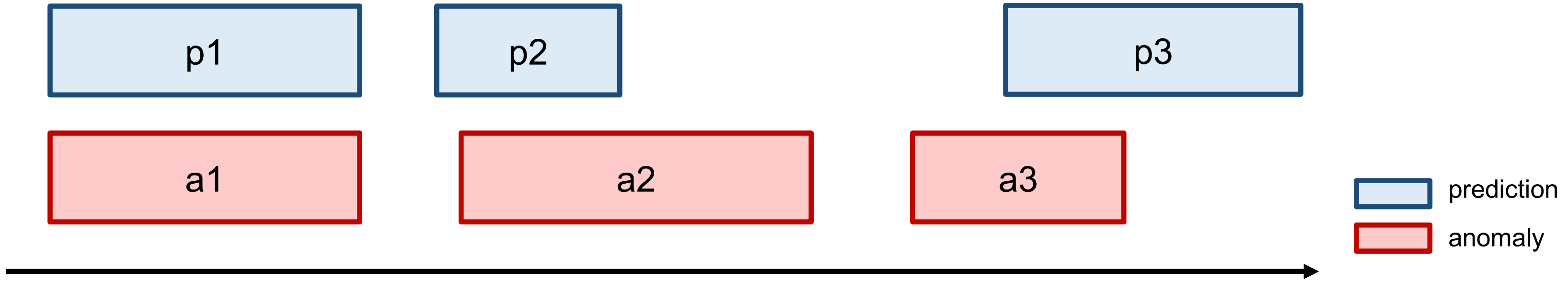
- **Detection method**
  - A detection method is used to identify anomalies in (time-series) data
  - A detection method is an assistance tool of anomaly detection

- **Accuracy metric**
  - Indicator of the detection method’s performance



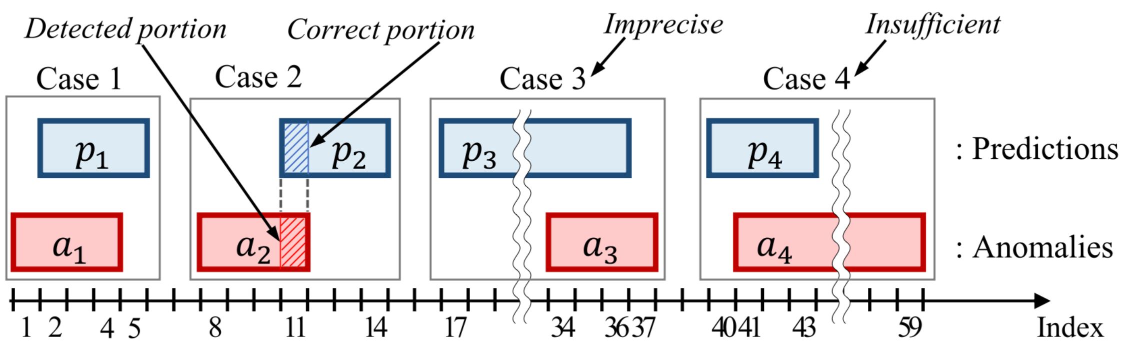
Motivation

- **Purpose**
  - Time-Series Anomaly Detection의 예측이 전문가에게 더 도움이 되는 것.(accuracy metric)

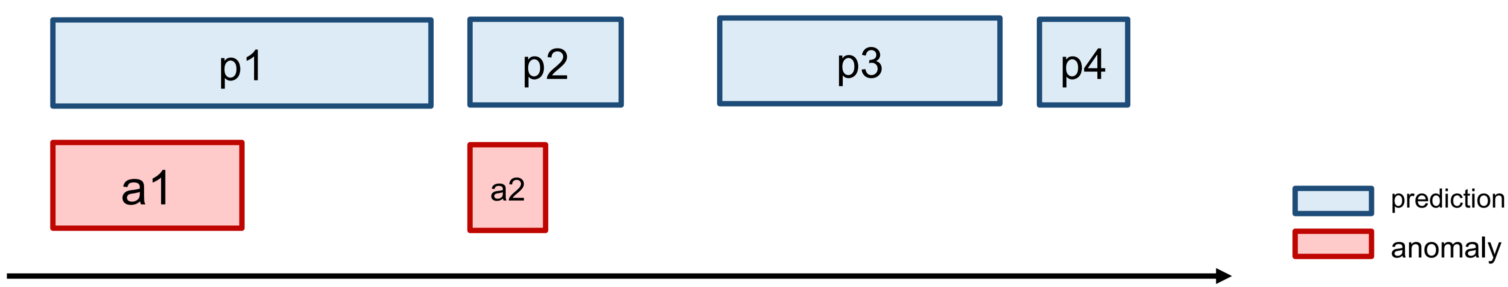


Motivation

- **Limitation of existing accuracy metrics**
  - overrate imprecise/insufficient cases



  - fail to penalize long incorrect prediction



Proposed Metrics: eTaPR

- **eTaR (recall like score)**
  - $eTaR = \frac{1}{|A|}\sum_{a \in A}\left(\frac{s^d(a)+s^d(a)\times s^p(a)}{2}\right)$
  - $s^d(a) = \begin{cases} 1, & if\ a \in \boxed{A^d} \\ 0, & otherwise \end{cases}$ $\quad s^p(a) = \frac{\sum_{p \in \boxed{P^c}} |a \cap p|}{|a|}$

- **eTaP(precision like score)**
  - $eTaP = \sum_{p \in P}\left(\frac{s^d(p)+s^d(p)\times s^p(p)}{2}\right)\times \boxed{w_p}$
  - $s^d(p) = \begin{cases} 1, & if\ p \in \boxed{P^c} \\ 0, & otherwise \end{cases}$ $\quad s^p(p) = \frac{\sum_{a \in \boxed{A^d}} |a \cap p|}{|p|}$ $\quad w_p = \frac{\sqrt{|p|}}{\sum_{q \in P} |q|}$

cross-referncing

weighting scheme

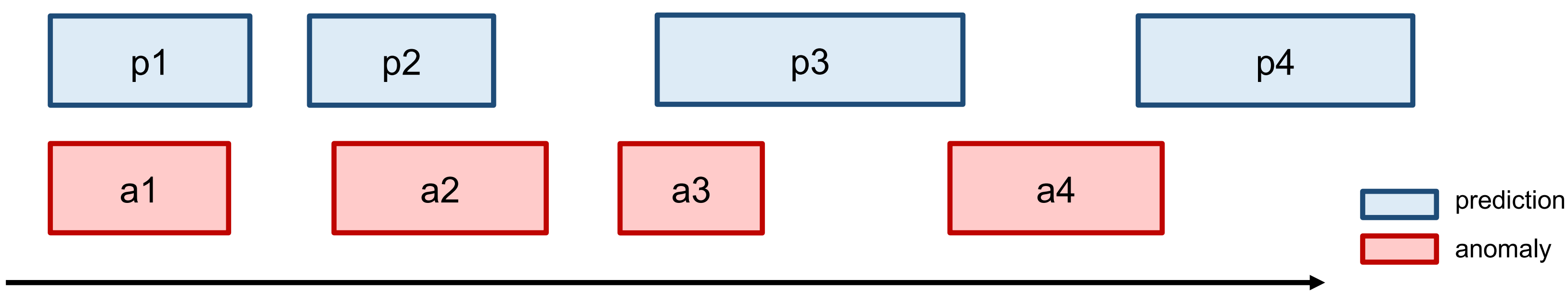
Proposed Metrics

- **Weighting scheme – long incorrect predicton**
  - prediction 길이를 가중치로 하는 w^p를 precision like score eTaP에 사용한다.
  - $w_p = \frac{\sqrt{|p|}}{\sum_{q \in P} \sqrt{|q|}}$

- **Cross-referencing – imprecise, insufficient**
  - A^d 집합과 P^c 집합이 서로 cross-reference해서 insufficient case, imprecise case를 filter한다.
  - A^d 집합과 P^c 집합을 eTaP, eTaR에 사용한다.
  - $A^d = \left\{ a \mid a \in A \ and \ \frac{\sum_{p \in P^c} |a \cap p|}{|a|} \geq \theta_r \right\}$
  - $P^c = \left\{ p \mid p \in P \ and \ \frac{\sum_{a \in A^d} |a \cap p|}{|p|} \geq \theta_p \right\}$

Proposed Metrics

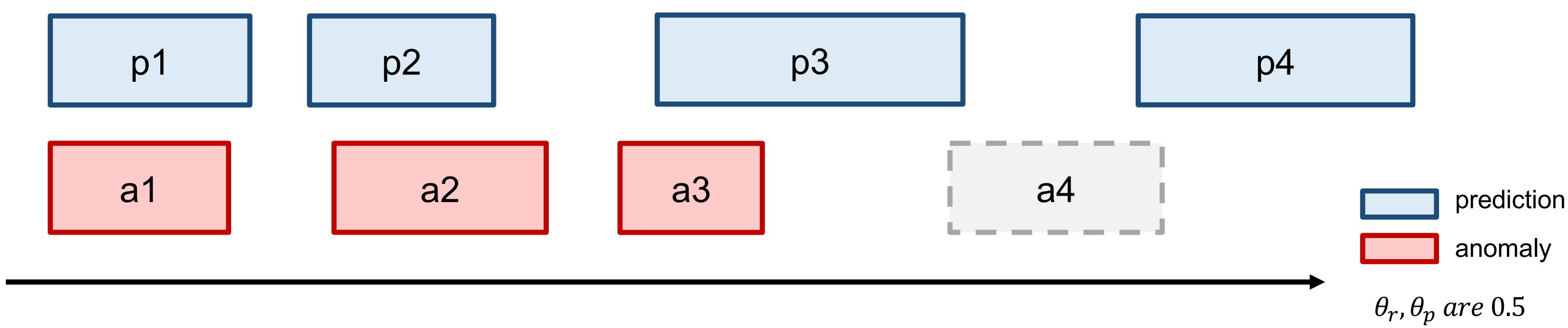
- **Cross-referencing – imprecise, insufficient**
  - A^d 집합과 P^c 집합이 서로 cross-reference해서 insufficient case, imprecise case를 filter한다.
  - $A^d = \left\{ a \mid a \in A \ and \ \frac{\sum_{p \in P^c} |a \cap p|}{|a|} \geq \theta_r \right\}$
  - $P^c = \left\{ p \mid p \in P \ and \ \frac{\sum_{a \in A^d} |a \cap p|}{|p|} \geq \theta_p \right\}$
  - P → $\boldsymbol{A^d}(1)$, $\boldsymbol{A^d(1)} \rightarrow \boldsymbol{P^c(1)}$
  - $\boldsymbol{P^c(1)} \rightarrow \boldsymbol{A^d(2)}$, $\boldsymbol{A^d(2)} \rightarrow \boldsymbol{P^c(2)}$, $\boldsymbol{P^c(2)} \rightarrow \boldsymbol{A^d(3)}$ … until $\boldsymbol{P^c(i)} = \boldsymbol{P^c(i-1)}$ and $\boldsymbol{A^d(i)} = \boldsymbol{A^d(i-1)}$



$\theta_r, \theta_p \ are \ 0.5$

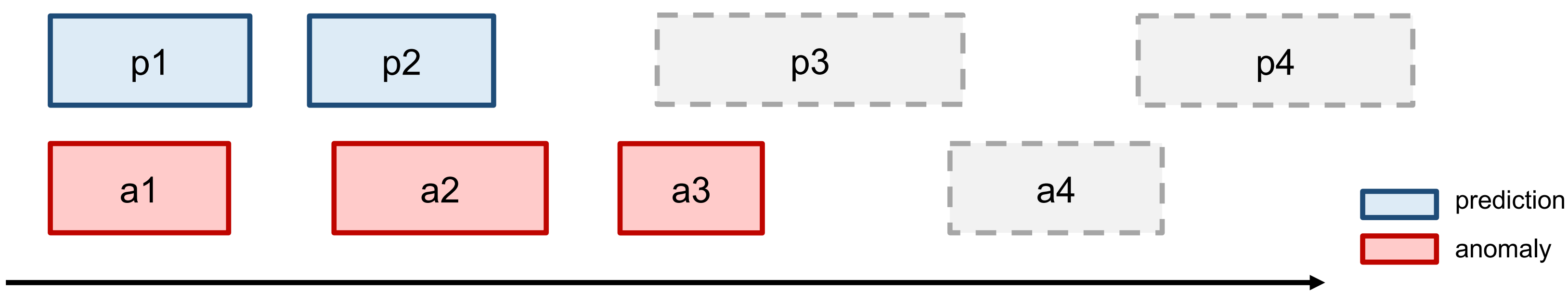
Proposed Metrics

- **Cross-referencing – imprecise, insufficient**
  - A^d 집합과 P^c 집합이 서로 cross-reference해서 insufficient case, imprecise case를 filter한다.
  - $A^d = \left\{ a \mid a \in A \ and \ \frac{\sum_{p \in P^c} |a \cap p|}{|a|} \geq \theta_r \right\}$
  - $P^c = \left\{ p \mid p \in P \ and \ \frac{\sum_{a \in A^d} |a \cap p|}{|p|} \geq \theta_p \right\}$
  - P $\rightarrow \boldsymbol{A^d}(1)$, $\boldsymbol{A^d}(\mathbf{1}) \rightarrow \boldsymbol{P^c}(\mathbf{1})$
  - $\boldsymbol{P^c}(\mathbf{1}) \rightarrow \boldsymbol{A^d}(\mathbf{2})$, $\boldsymbol{A^d}(\mathbf{2}) \rightarrow \boldsymbol{P^c}(\mathbf{2})$, $\boldsymbol{P^c}(\mathbf{2}) \rightarrow \boldsymbol{A^d}(\mathbf{3})$ … until $\boldsymbol{P^c}(\boldsymbol{i}) = \boldsymbol{P^c}(\boldsymbol{i}-\mathbf{1})$ and $\boldsymbol{A^d}(\boldsymbol{i}) = \boldsymbol{A^d}(\boldsymbol{i}-\mathbf{1})$



Proposed Metrics

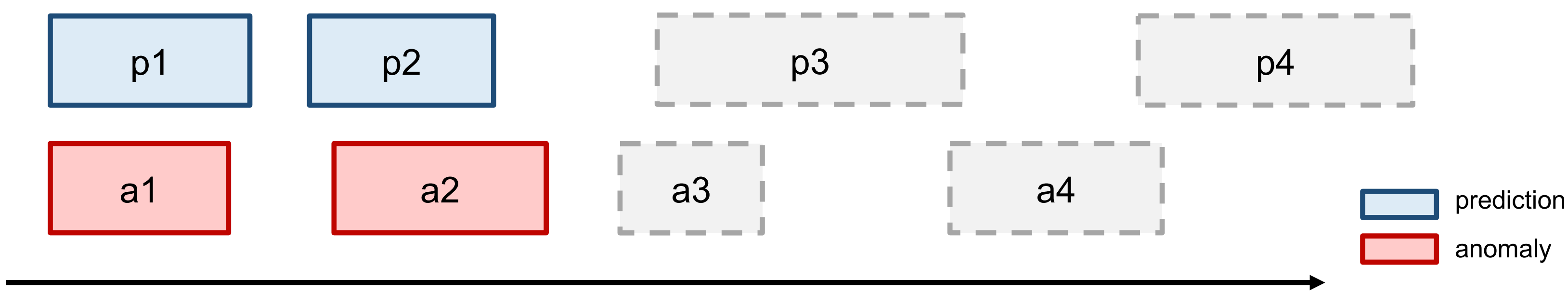
- **Cross-referencing – imprecise, insufficient**
    - A^d 집합과 P^c 집합이 서로 cross-reference해서 insufficient case, imprecise case를 filter한다.
    - $A^d = \left\{ a \mid a \in A \ and \ \frac{\sum_{p \in P^c} |a \cap p|}{|a|} \geq \theta_r \right\}$
    - $P^c = \left\{ p \mid p \in P \ and \ \frac{\sum_{a \in A^d} |a \cap p|}{|p|} \geq \theta_p \right\}$
    - P $\rightarrow \boldsymbol{A^d}(1)$, $\boldsymbol{A^d(1) \rightarrow P^c(1)}$
    - $\boldsymbol{P^c(1) \rightarrow A^d(2)}$, $\boldsymbol{A^d(2) \rightarrow P^c(2)}$, $\boldsymbol{P^c(2) \rightarrow A^d(3)}$ … until $\boldsymbol{P^c(i) = P^c(i-1)}$ and $\boldsymbol{A^d(i) = A^d(i-1)}$



$\theta_r, \theta_p \ are \ 0.5$

Proposed Metrics

- **Cross-referencing – imprecise, insufficient**
    - A^d 집합과 P^c 집합이 서로 cross-reference해서 insufficient case, imprecise case를 filter한다.
    - $A^d = \left\{a \mid a \in A \; and \; \frac{\sum_{p \in P^c} |a \cap p|}{|a|} \geq \theta_r \right\}$
    - $P^c = \left\{p \mid p \in P \; and \; \frac{\sum_{a \in A^d} |a \cap p|}{|p|} \geq \theta_p \right\}$
    - P $\rightarrow \boldsymbol{A^d}(1)$, $\boldsymbol{A^d}(\mathbf{1}) \rightarrow \boldsymbol{P^c}(\mathbf{1})$
    - $\boldsymbol{P^c}(\mathbf{1}) \rightarrow \boldsymbol{A^d}(\mathbf{2})$, $\boldsymbol{A^d}(\mathbf{2}) \rightarrow \boldsymbol{P^c}(\mathbf{2})$, $\boldsymbol{P^c}(\mathbf{2}) \rightarrow \boldsymbol{A^d}(\mathbf{3})$ … until $\boldsymbol{P^c}(\boldsymbol{i}) = \boldsymbol{P^c}(\boldsymbol{i}-\mathbf{1})$ and $\boldsymbol{A^d}(\boldsymbol{i}) = \boldsymbol{A^d}(\boldsymbol{i}-\mathbf{1})$



θ_r, θ_p *are* 0.5

Proposed Metrics: eTaPR

- **eTaR (recall like score)**
  - $eTaR = \frac{1}{|A|}\sum_{a\in A}\left(\frac{s^d(a)+s^d(a)\times s^p(a)}{2}\right)$
  - $s^d(a) = \begin{cases} 1, & if\ a \in \boxed{A^d} \\ 0, & otherwise \end{cases}$ $\quad s^p(a) = \frac{\sum_{p\in \boxed{P^c}} |a \cap p|}{|a|}$

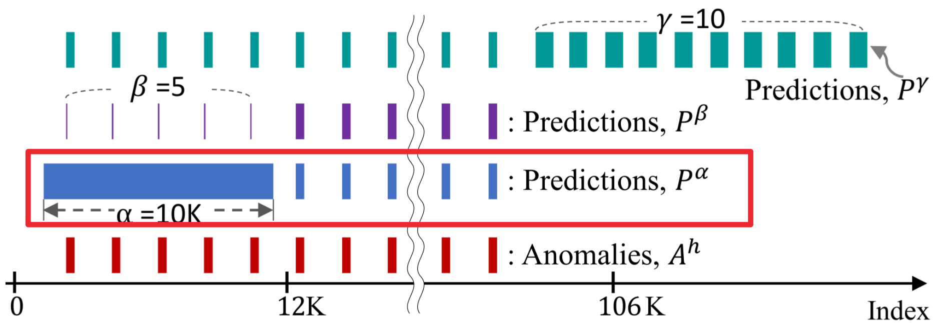
- **eTaP(precision like score)**
  - $eTaP = \sum_{p\in P}\left(\frac{s^d(p)+s^d(p)\times s^p(p)}{2}\right) \times \boxed{w_p}$
  - $s^d(p) = \begin{cases} 1, & if\ p \in \boxed{P^c} \\ 0, & otherwise \end{cases}$ $\quad s^p(p) = \frac{\sum_{a\in \boxed{A^d}} |a \cap p|}{|p|}$ $\quad w_p = \frac{\sqrt{|p|}}{\sum_{q\in P}|q|}$

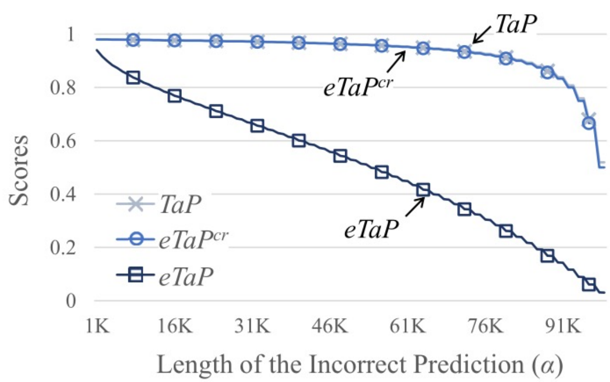
cross-referencing

weighting scheme

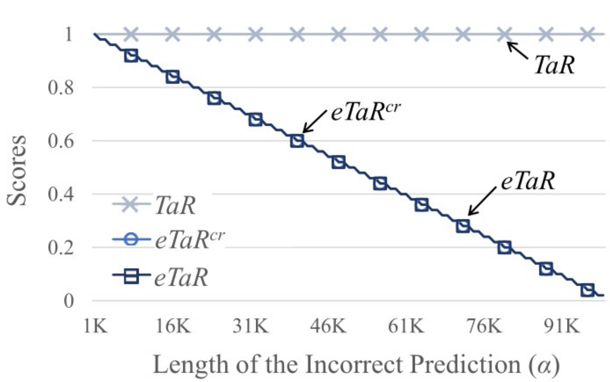
Experimental Results

- **Hypothetical datasets**
  - Imprecise cases (also lengthy incorrect prediction)





(a) Precision-like scores



(b) Recall-like scores

Experimental Results

- **Hypothetical datasets**
  - Imprecise cases (also lengthy incorrect prediction)

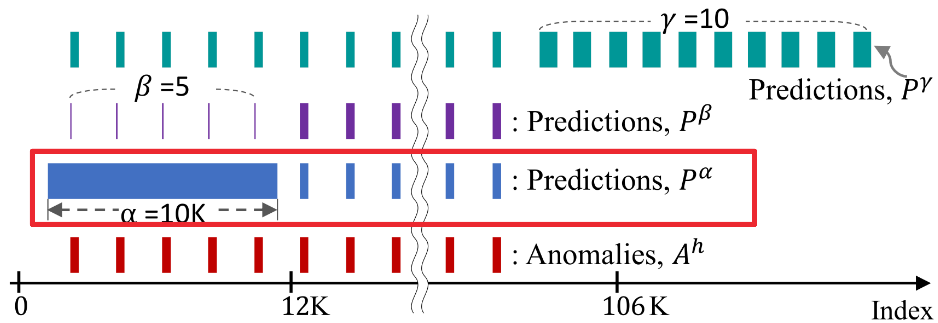
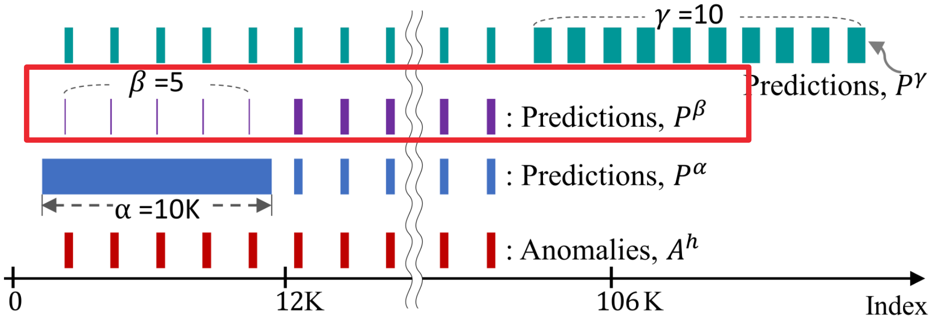


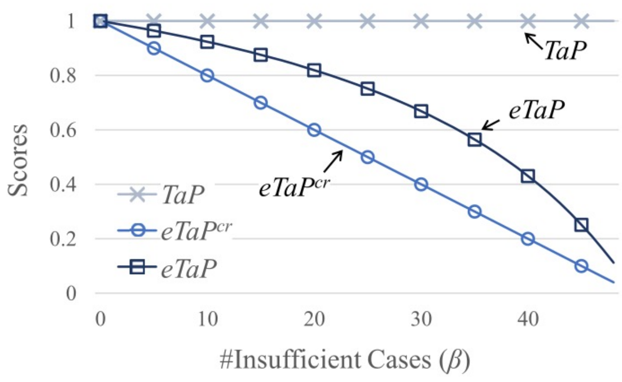
Table 2: Evaluations on the first hypothetical dataset (i.e., A^h and P^{α})

Pred.	PR			PA-PR			TSAD			TaPR			eTaPR		
	Prec.	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*	*RbP*	*RbR*	*F1*	*TaP*	*TaR*	*F1*	*eTaP*	*eTaR*	*F1*
$P^{\alpha=25K}$	0.18	1.00	0.31	0.18	1.00	0.31	0.97	1.00	0.98	0.97	1.00	0.98	0.70	0.74	0.72
$P^{\alpha=50K}$	0.10	1.00	0.18	0.10	1.00	0.18	0.96	1.00	0.98	0.96	1.00	0.98	0.53	0.50	0.51
$P^{\alpha=75K}$	0.07	1.00	0.13	0.07	1.00	0.13	0.92	1.00	0.96	0.93	1.00	0.96	0.31	0.24	0.27
$P^{\alpha=100K}$	0.05	1.00	0.10	0.05	1.00	0.10	0.00	1.00	0.00	0.03	1.00	0.06	0.00	0.00	-

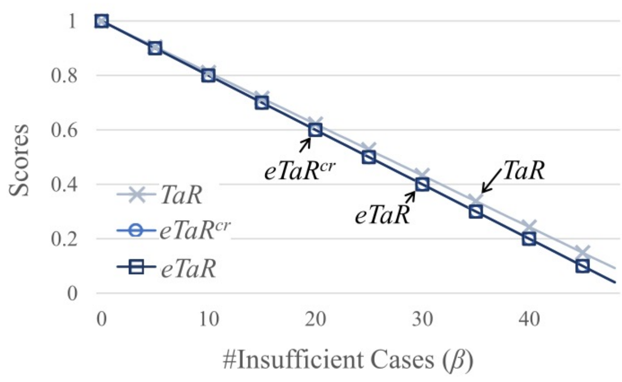
Experimental Results

- **Hypothetical datasets**
  - insufficient cases





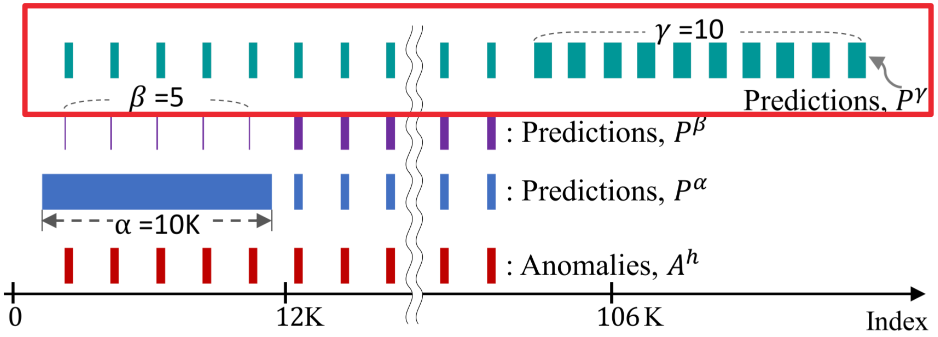
(a) Precision-like scores

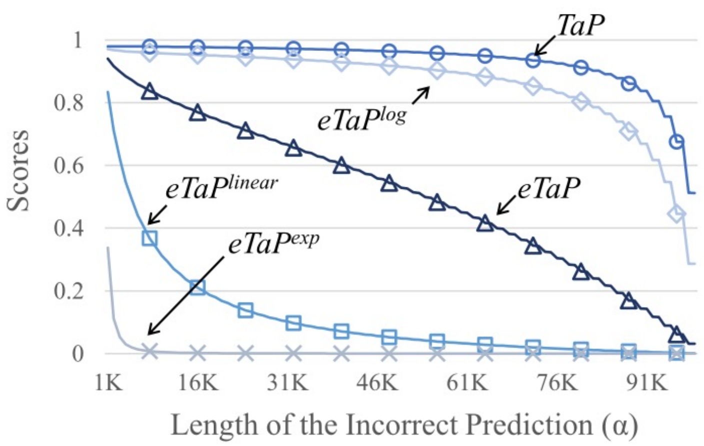


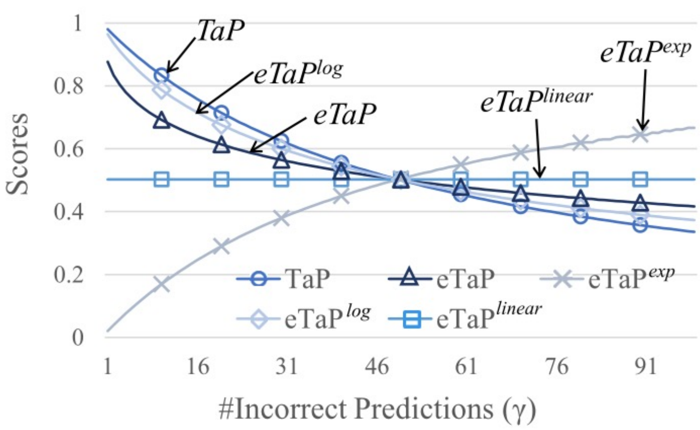
(b) Recall-like scores

Experimental Results

- **Hypothetical datasets**
  - incorrect predictions(length and frequency)

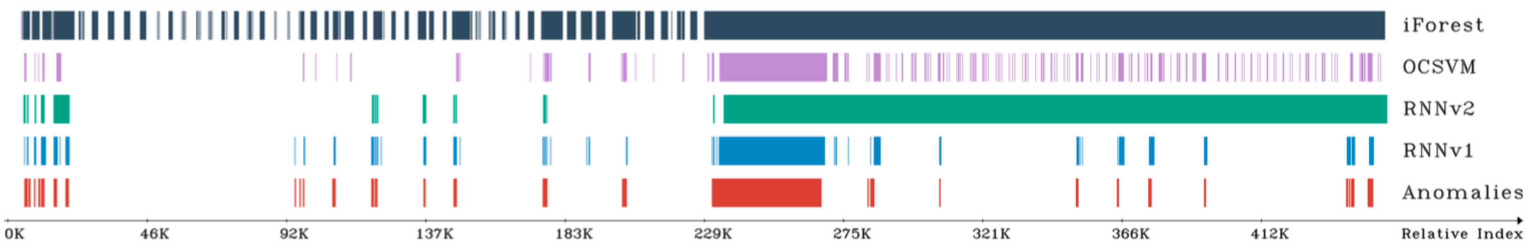






Experimental Results

- **SWaT dataset**



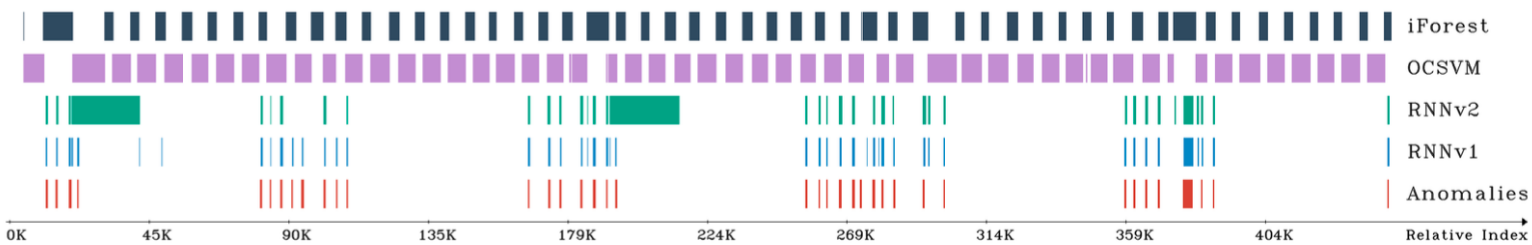
(a) Predictions and anomalies on SWaT dataset.

Table 3: Evaluations for the detection methods on the SWaT dataset (Bold indicates the best F1 score in each metric.)

Pred.		**PR**			**PA-PR**			**TSAD**			**TaPR**			**eTaPR**	
	Prec.	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*	*RbP*	*RbR*	*F1*	*TaP*	*TaR*	*F1*	*eTaP*	*eTaR*	*F1*
RNNv1	0.81	0.78	**0.79**	0.84	0.95	**0.89**	0.49	0.41	0.45	0.68	0.53	0.60	0.59	0.55	**0.57**
RNNv2	0.19	0.80	0.31	0.21	0.92	0.34	0.55	0.59	**0.57**	0.83	0.63	**0.72**	0.17	0.20	0.18
OCSVM	0.65	0.75	*0.70*	0.68	0.85	*0.76*	0.07	0.27	0.11	0.07	0.30	0.11	0.23	0.28	0.25
iForest	0.17	0.95	0.29	0.18	0.99	0.30	0.04	0.68	0.08	0.05	0.82	0.09	0.07	0.12	0.09

Experimental Results

- **HAI dataset**



(b) Predictions and anomalies on HAI dataset.

Table 4: Evaluations for the detection methods on the HAI datase (Bold indicates the best F1 score in each metric.)

Pred.	**PR**			**PA-PR**			**TSAD**			**TaPR**			**eTaPR**		
	Prec.	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*	*RbP*	*RbR*	*F1*	*TaP*	*TaR*	*F1*	*eTaP*	*eTaR*	*F1*
RNNv1	0.80	0.73	**0.76**	0.85	0.95	**0.90**	0.74	0.35	0.48	0.84	0.82	**0.83**	0.77	0.80	**0.78**
RNNv2	0.23	0.84	0.36	0.25	0.93	0.39	0.56	0.79	**0.66**	0.73	0.84	0.78	0.51	0.72	0.60
OCSVM	0.03	0.49	0.06	0.03	0.52	0.06	0.03	0.50	0.06	0.05	0.57	0.09	0.00	0.01	-
iForest	0.06	0.51	0.11	0.06	0.55	0.11	0.03	0.40	0.06	0.05	0.44	0.09	0.00	0.01	-

Conclusion

- **Time-Series Anomaly Detection : Expert Scenario를 생각해야한다.**
- **cross-referencing (eTaPR)**
  - imprecise, insufficient prediction, anomaly에 점수 주는 것을 방지
- **weighting scheme (eTaP)**
  - penalize lengthy incorrect prediction