国立大学法人 静岡大学

データストリーム上の突発イベントを検知するオンラインアルゴリズムの検討

○山本 泰生, Thanapol Phungtua-eng

静岡大学・情報学部

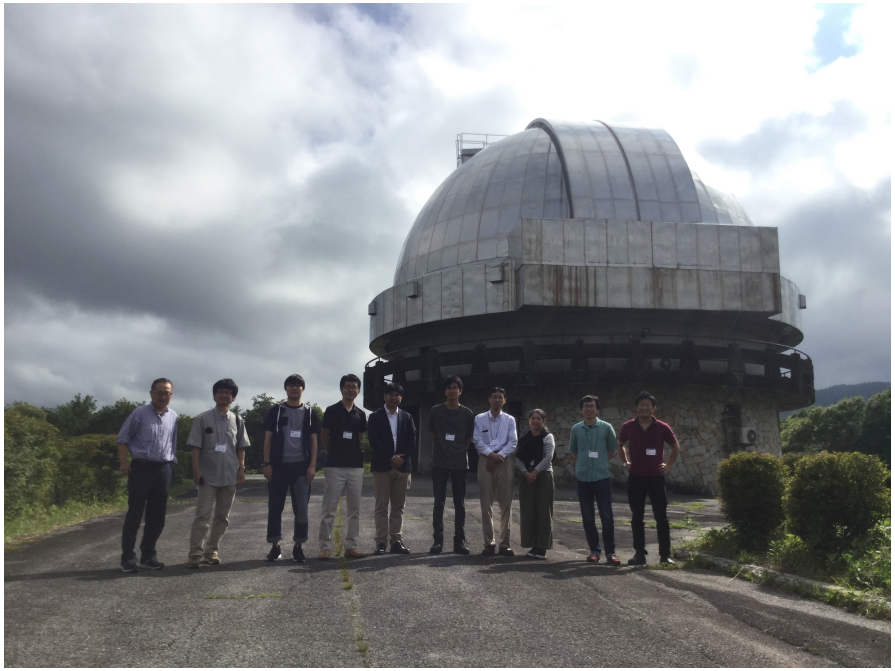
木曽シュミットシンポジウム2021

2021年10月6日 (オンライン発表)

発表の流れ

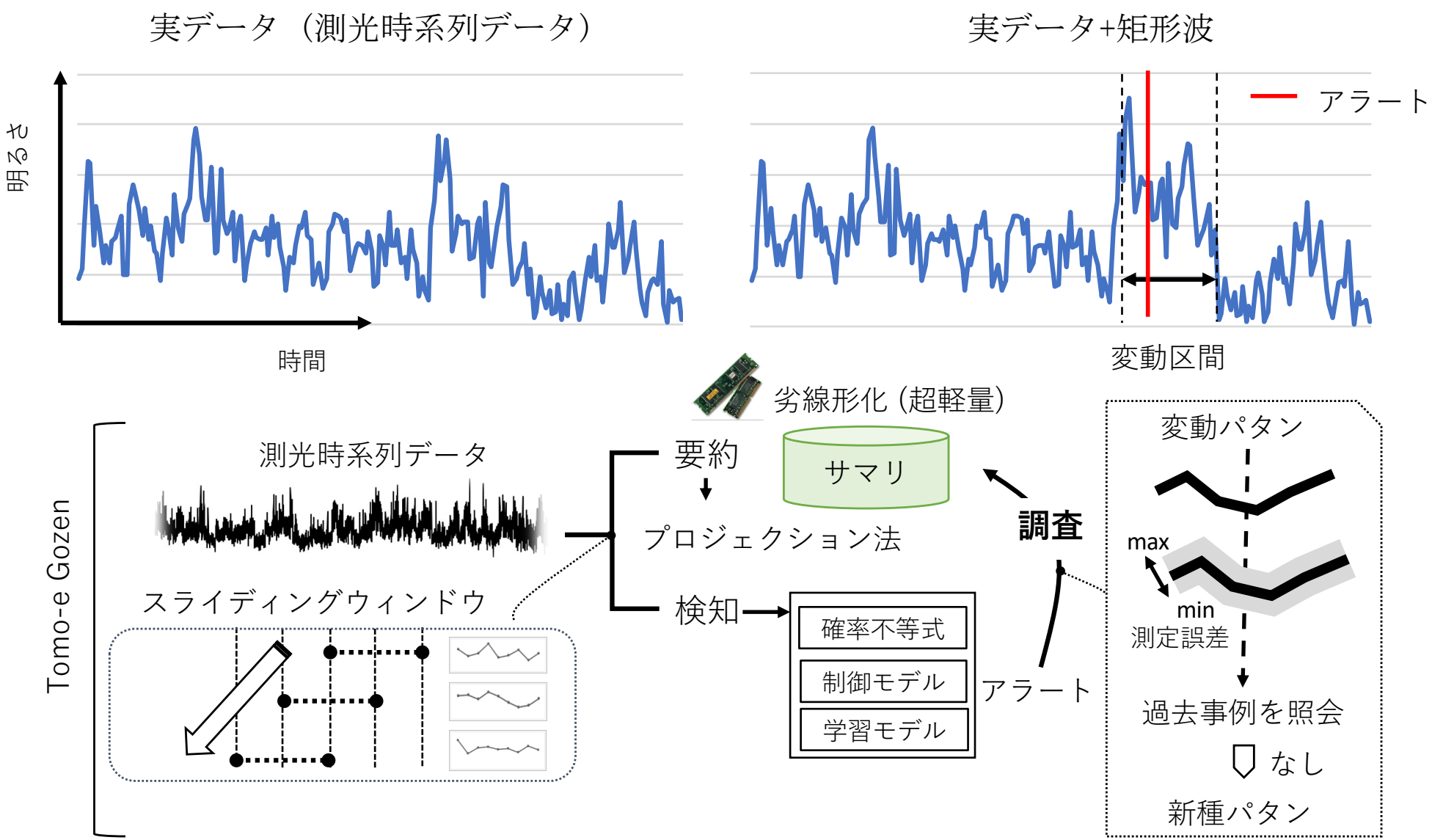
- 背景
- アプローチ
- 関連研究
- 提案法
- 実験
- まとめ

今回で２度目
「木曽シュミットシンポジウム2019」から参加



背景: 課題の概要

- 未知の突発的天体現象をリアルタイム検知する問題



背景: 異常検知法の種類

- 事例に基づく手法
  - 最近傍法
  - クラスタリング
- モデルに基づく手法
  - 制御モデル (カルマンフィルタ等)
  - 時系列モデル (HMM, Bayesian, RNN, TCN)
  - 確率分布モデル (ガウス過程)
- 圧縮に基づく手法
  - 次元縮退 (PCA, SVD, 圧縮センシング)
  - 簡潔データ構造 (Random Cut Forest)

背景: 異常検知問題としての特徴 (1)

S/N が低い

一般的な問題設定

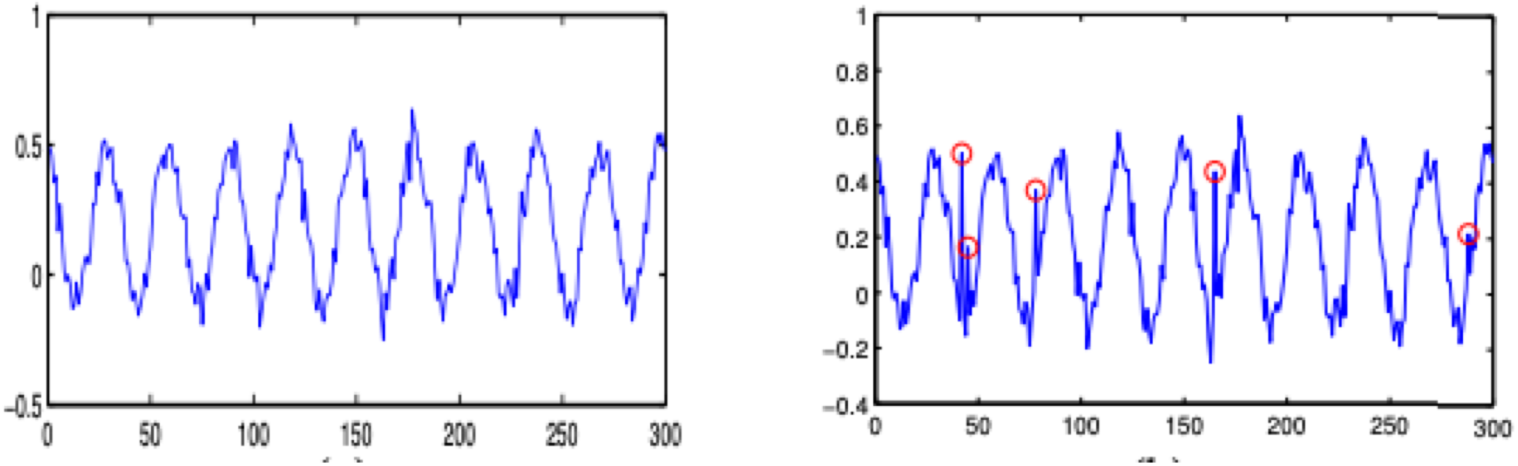
パターン中に出現するノイズ (異常) を見つける

今回の問題設定

ノイズ中に出現するパターン (異常) を見つける

👉 「ノイズの砂漠からダイヤを見つける」難しさ

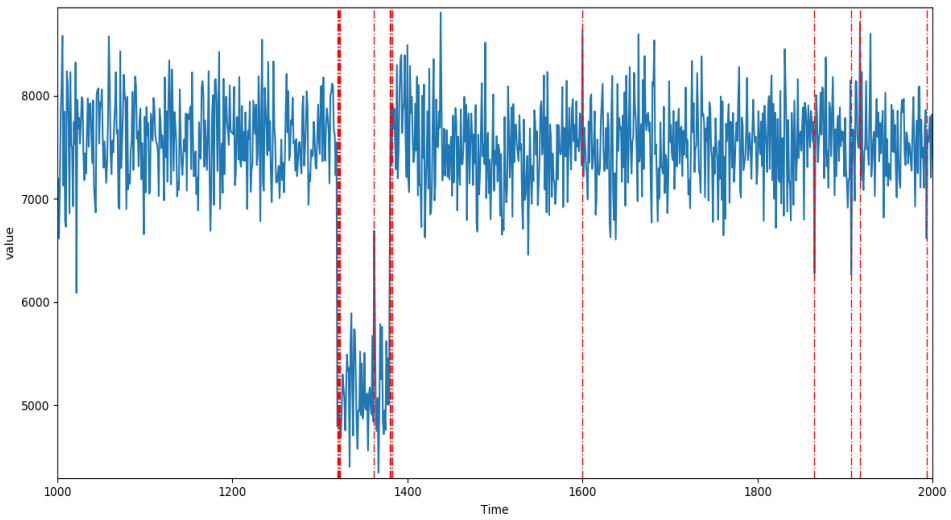
センサーネットワークトラフィックの異常検知の例



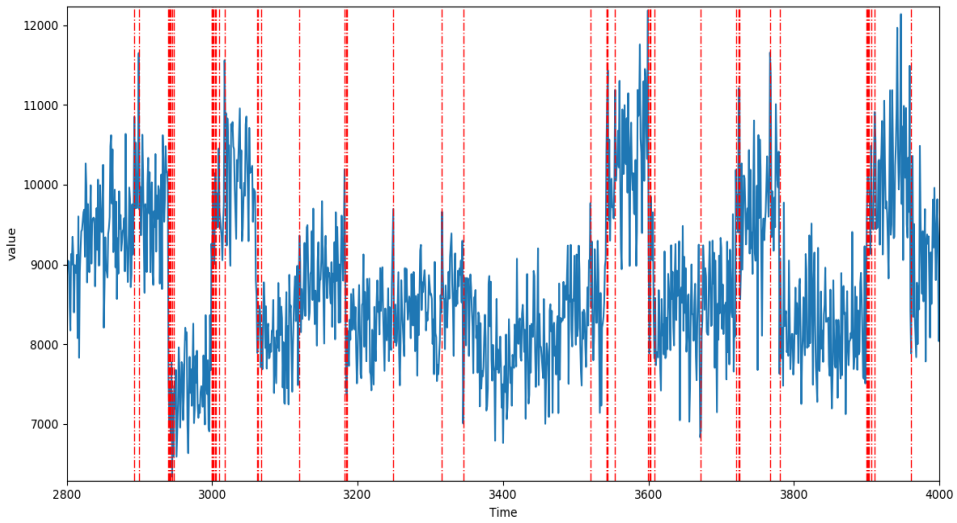
背景: 異常検知問題としての特徴 (2)

どんなパターンがいつ，なぜ発生するか不明

- 形状, 時間幅, 強度
- 頻度
- 原因 (天体由来/機器由来/環境由来)



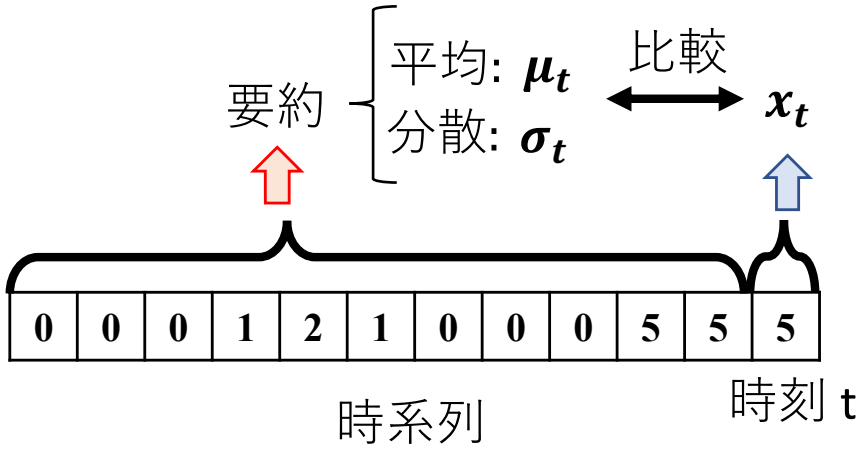
Gaia 4451658397468928００ PCA



Gaia 4124250738279360０ PCA

アプローチ: 確率不等式に基づく検知

- 特徴
  - 集中不等式に基づき異常値を検出する
    (+) 観測値が互いに独立な系列に有効
    (+) 高速・軽量な検知が可能
    (+) 生成モデルに依存せず利用可能
- アイデア
  - 正常系の要約統計量 (平均 $\boldsymbol{\mu_t}$, 分散 $\boldsymbol{\sigma_t}$) をもとに
    各時刻の値 (x_t) の異常判定を行う

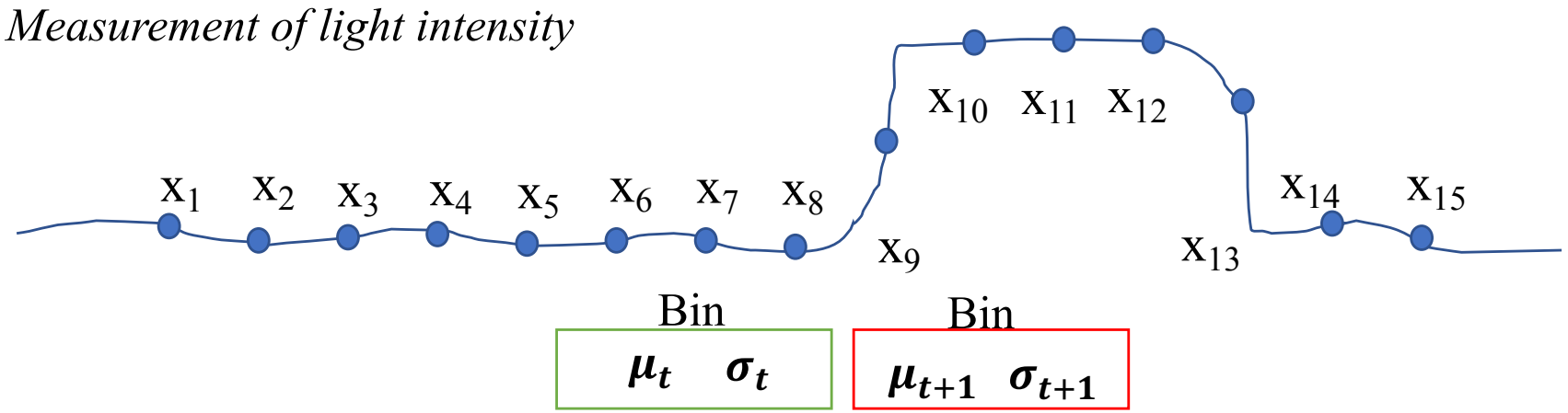


$$P(|x_t - \mu_t| \geq k\sigma_t) \leq \frac{1}{k^2}$$

Chebyshev's inequality

先行研究: SK法

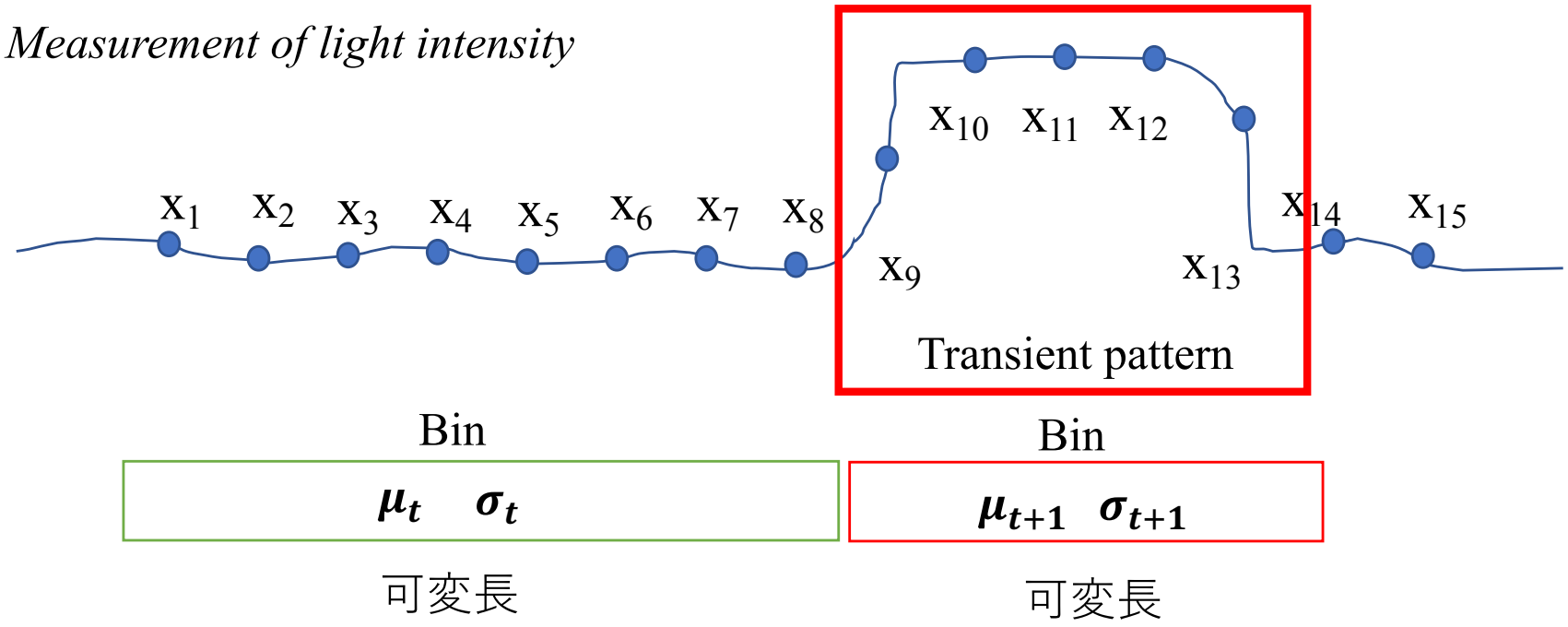
- ビニングによる要約 G. Shevlyakov and M. Kan, *FRUCT* 2020
  - 固定長のビニングを用いて隣接する Bin の統計量を比較する
    (+) シンプルで高速な判定が可能
    月レーザー測距データの異常検知の利用
    (-) 突発イベントの duration によって検知性能が変動する



G. Shevlyakov and M. Kan, "Stream Data Preprocessing: Outlier Detection Based on the Chebyshev Inequality with Applications," *2020 26th Conference of Open Innovations Association (FRUCT)*, 2020

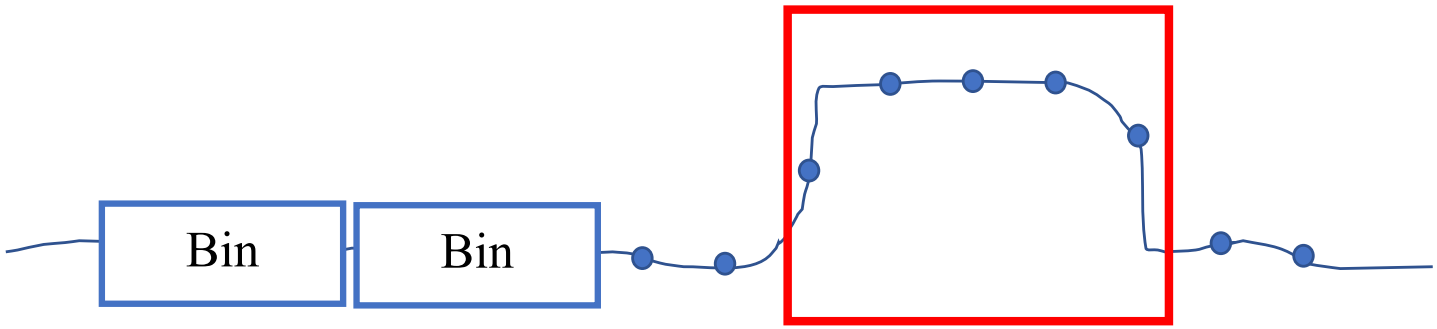
提案手法: 動的ビニング

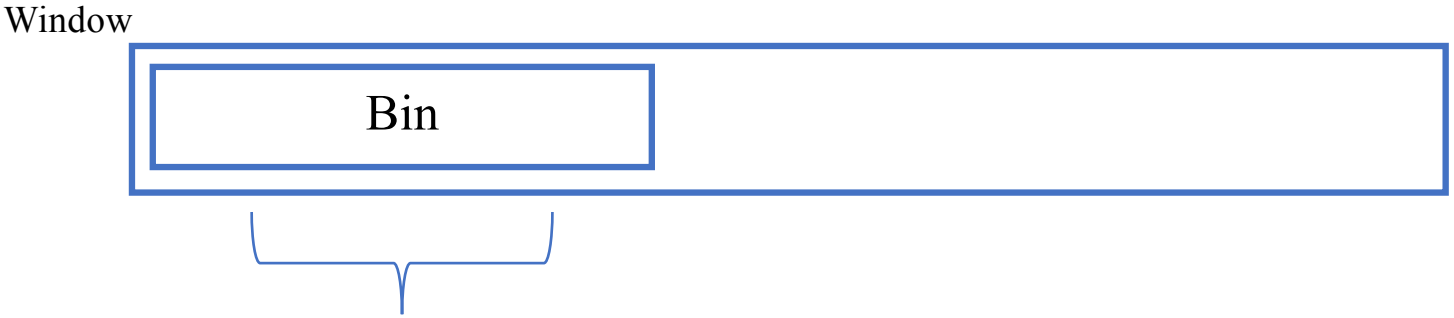
- アイデア: ビニング幅を Auto-Focus する
  - ➢ 類似する隣接 Bin はマージして一つにする



T. Phungtua-eng, Y. Yamamoto and S. Sako, ”Detection for transient patterns with unpredictable duration using Chebyshev Inequality and dynamic”,
2021 Eighth International Symposium on Computing and Networking Workshops (CANDARW), 2021 to appear

Dynamic binning





T. Phungtua-eng, Y. Yamamoto and S. Sako, ”Detection for transient patterns with unpredictable duration using Chebyshev Inequality and dynamic”,
2021 Eighth International Symposium on Computing and Networking Workshops (CANDARW), 2021 to appear

Dynamic binning

Algorithm 1: *DynamicBin*

input: W: window, α: significance level

1 **Compute** T-test value T_{test} of Equation 1. from Bin_l and Bin_{l-1}
2 **if** $|T_{test}| \leq T_{1-\frac{\alpha}{2}}$ **then**
3 **Compute** F-test value F_{test} of Equation 2 from Bin_l and Bin_{l-1}
4 **if** $F_{test} \leq F_{1-\frac{\alpha}{2}}$ **then**
5 **Merge**(Bin_l, Bin_{l-1})
6 **else**
7 Search the bin Bin_i whose T-test value is minimum in W
8 **Merge**(Bin_i, Bin_{i-1}).
9 **end**
10 **end**

- Line 2 : Test two bins are equal mean.
  - $$\text{T-test} = \frac{\mu_i - \mu_{i-1}}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_{i-1}^2}{n_{i-1}}}} \quad (1)$$
- Line 4 : Test two bins are equal maximum variance.
  - $$\text{F-test} = \frac{\sigma^2 max_1}{\sigma^2 max_2} \quad (2)$$
- Line 7 : If two bins are not likely to be similar.
  - We search the Bin whose the least T-test in the window.

T. Phungtua-eng, Y. Yamamoto and S. Sako, ”Detection for transient patterns with unpredictable duration using Chebyshev Inequality and dynamic”,
2021 Eighth International Symposium on Computing and Networking Workshops (CANDARW), 2021 to appear

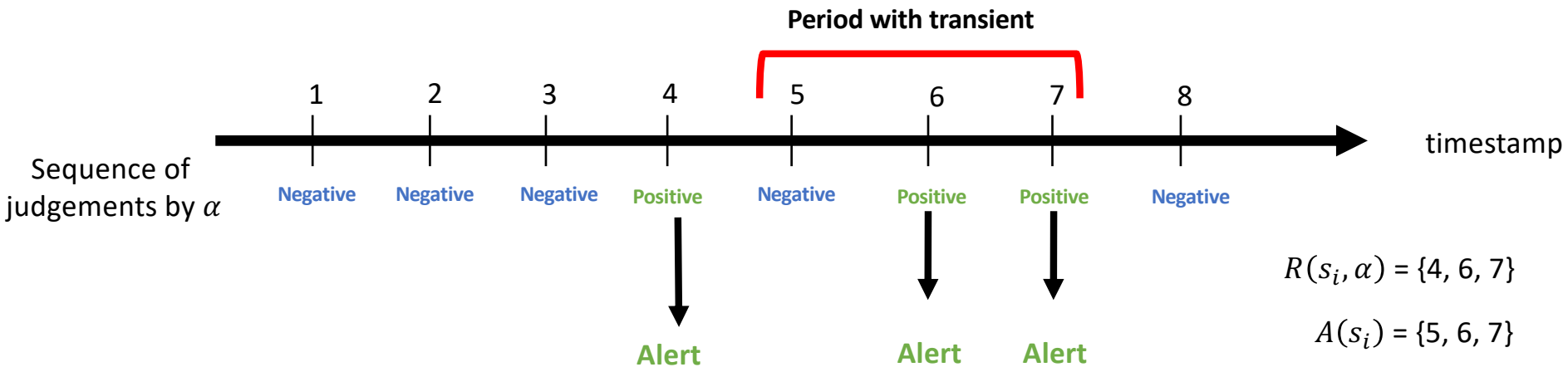
実験設定その１：ベンチマークデータ

- 5000天体のライトカーブデータ (逢澤様よりご提供)
  - 背景光除去データ (bgs)
  - 大気ゆらぎ除去データ (pca)
  - 対象: 3000サンプル以上&平均値5000~10000の天体
    - ✓ 訓練用: 108天体
    - ✓ **テスト用: 476天体**
  - 突発パターンとして**矩形波**を追加
    - ✓ 区間幅: **10サンプル**, **30サンプル**, 60サンプル
    - ✓ 振幅: 3σ, 5σ

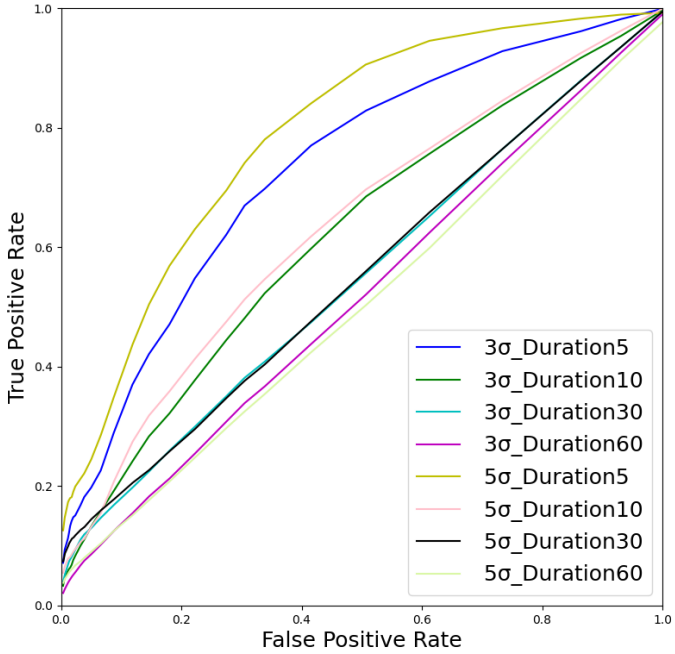
実験設定その2: 評価方法

Notion

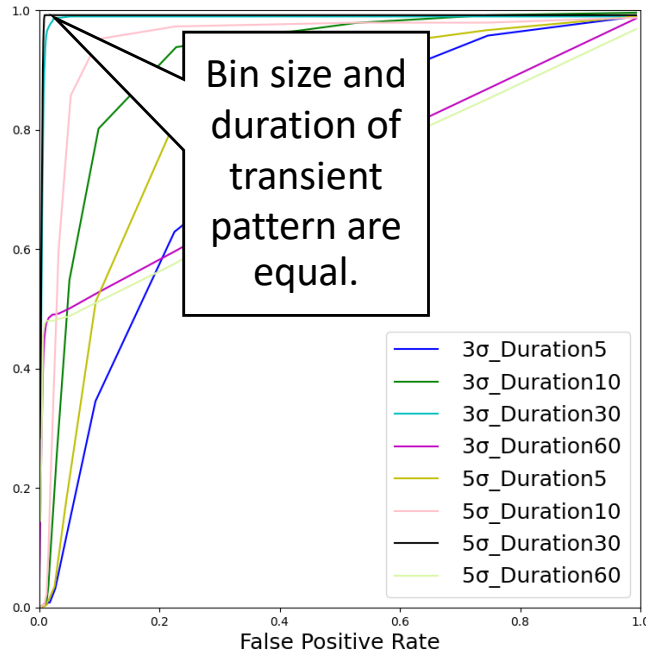
- α: detection method
- S: $\{s_1, s_2, \cdots, s_n\}$ where s_i is the light-curve of the i_{th} star
- $R(s_i, \alpha)$: set of timestamps in s_i, each of when α publishes an alert
- $A(s_i)$: set of timestamps for the period with a transient pattern



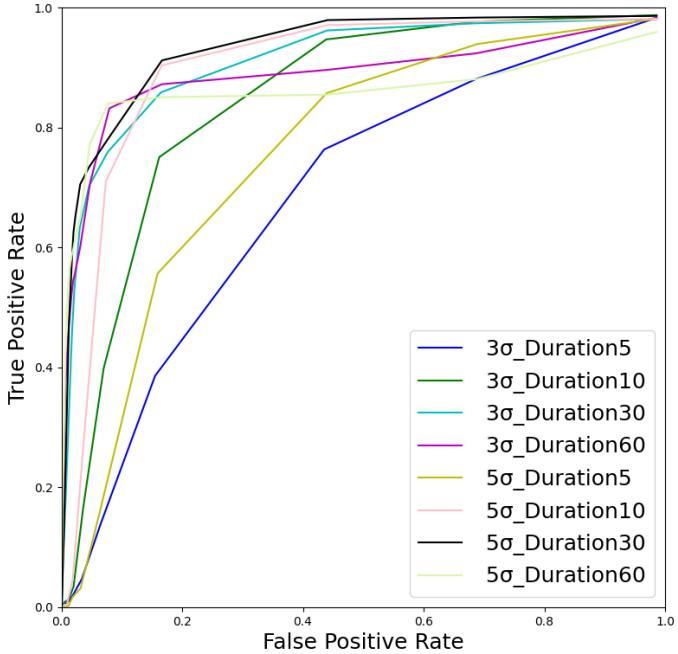
Influence of bin size (SK method)



SK method
Bin Size= 3

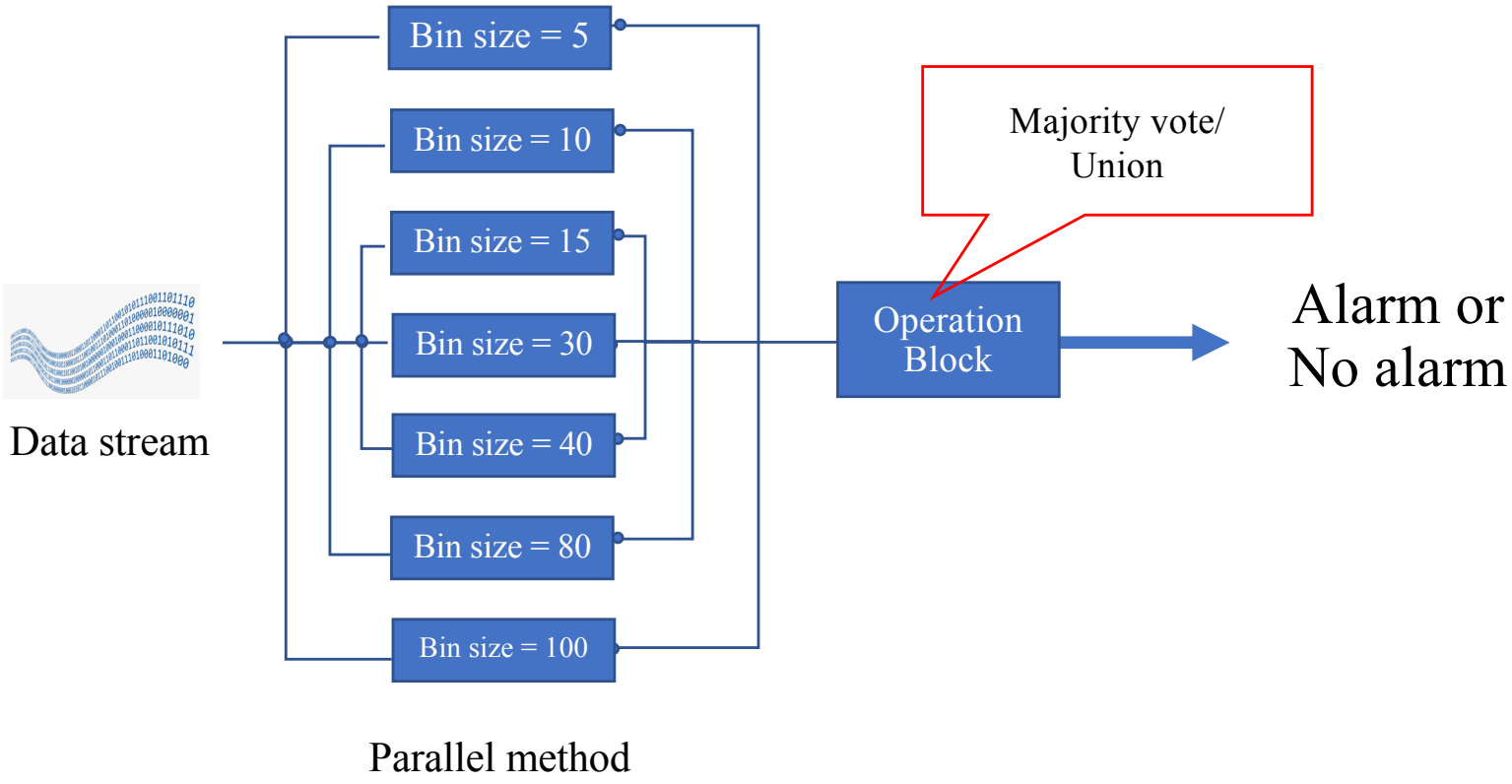


SK method
Bin Size= 30

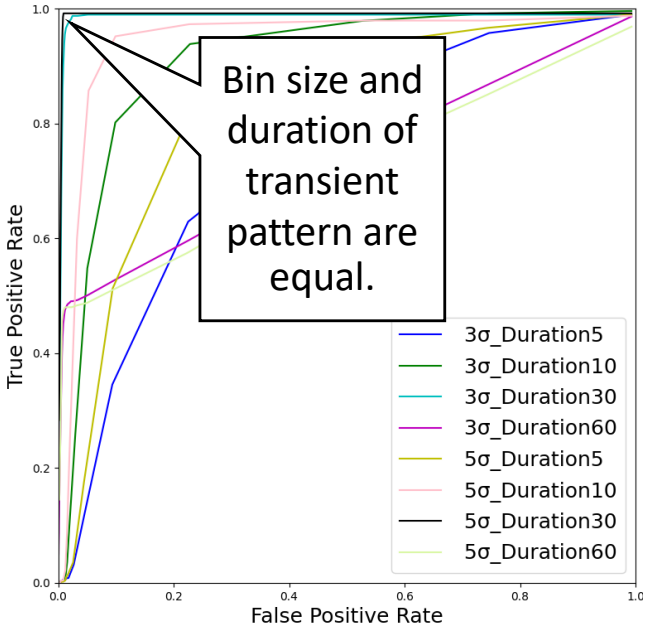


SK method
Bin Size= 50

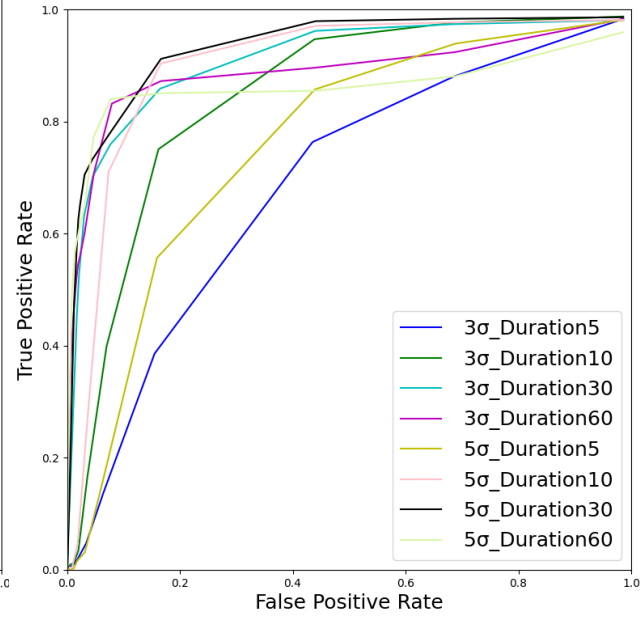
Parallel method of SK method



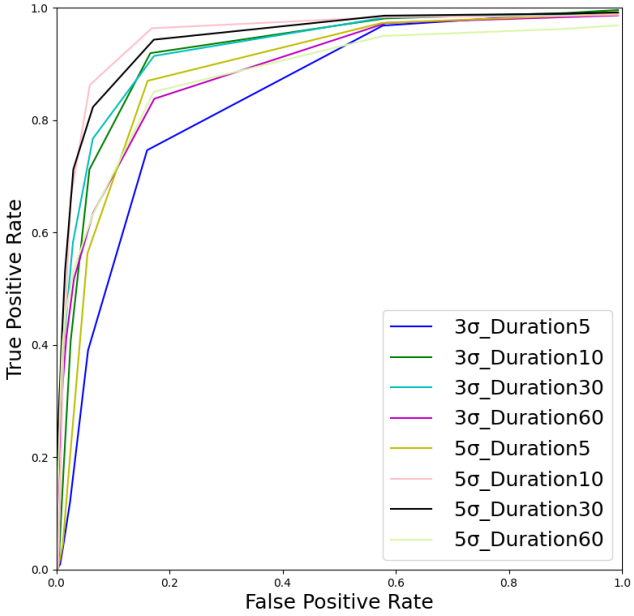
Parallel method : Majority vote



SK method
Bin Size= 30



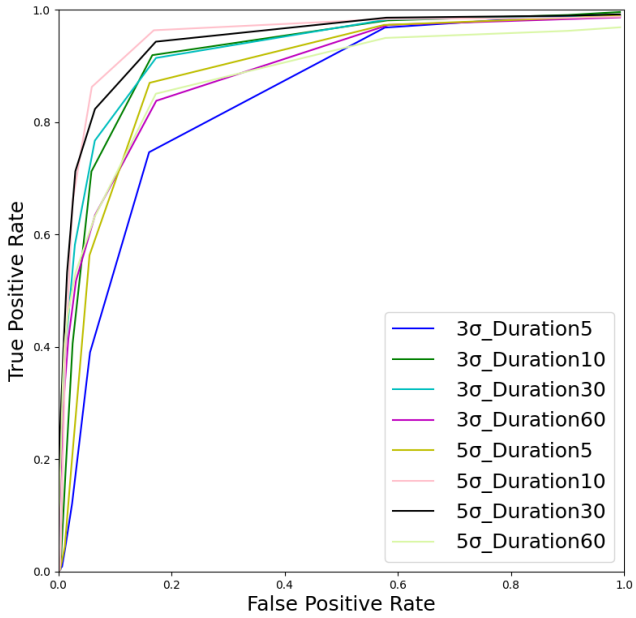
SK method
Bin Size= 50



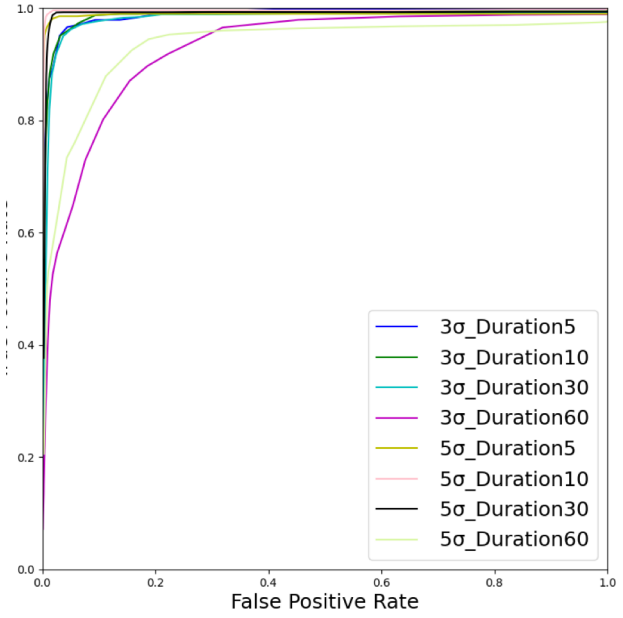
Majority vote
Bin size = [5, 10, 15, 30, 40, 80, 100]

Result of Dynamic binning

先行研究

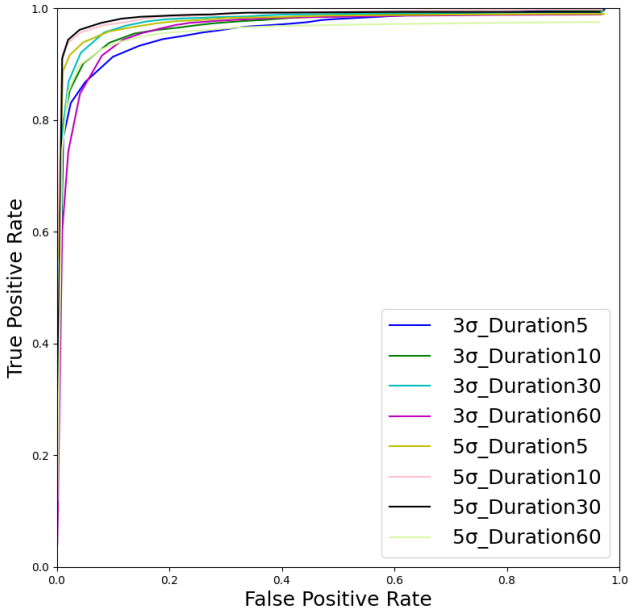


Majority method
[5, 10, 30, 40, 80, 100]
(SK method)



Union with bin
[5, 10, 30, 40, 80, 100]
(SK method)

提案法



Dynamic Binning

まとめと今後の課題

- 確率不等式に基づく手法の紹介
- ベンチマークデータの整備・拡張
  - ➢ 元データに出現している突発パターンの確認
  - ➢ パターン形状・頻度のバリエーションを増やす
- 手法の改良
  - ➢ 確率不等式ベース検知法の高次 (多段) 化
    - 👉 任意の時間幅のパターン検知を目指す
  - ➢ 劣線形サマリに基づく異常検知法の開発
    - 👉 カーネル密度推定に基づくサマリを利用する
  - ➢ 圧縮に基づく既存法の性能評価
    - 👉 RCF, SVD+圧縮センシング