# Online Multiscale Method for Change Detection in Automated Data-Quality Monitoring

Ronglong Fang [1]
in collaboration with Abdullah Farhat [1] Yuesheng Xu [1] Markus Diefenthaler[2] Holly Szumila-Vance[2]

[1]Old dominion university, Norfolk, VA, USA.

[2]Jefferson lab, Newport news, VA, USA

Artificial Intelligence for Science, Industry and Society, October 11-15, 2021

## Overview

1. Automated Data-Quality Monitoring
2. Online Multiscale method
   - Change detect for continuous function
   - A simple example
   - Change detection for streaming data
3. Results for Physics Data

## Automated Data-Quality Monitoring

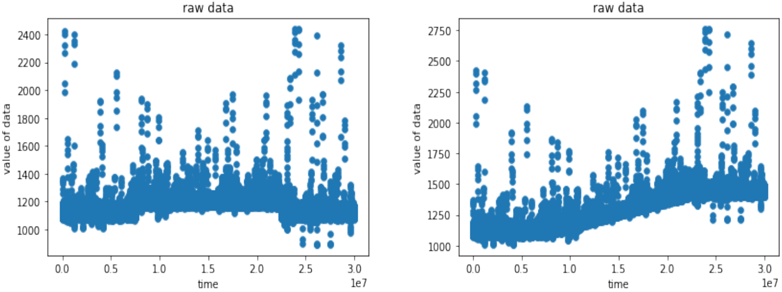
In most challenging data analysis applications, data evolve over time and must be analyzed in near real time [3].

In experimental physics, it possible takes a year or longer to obtain data. If we develop online machine learning method to monitor data, we can find problems in data while data taking. We can make the detector more stable and make data more believable.

To deal with time-changing data, our goal is :

1. detect when a change occurs.
2. detect what kind of change occurs, e.g., sudden change occurs or linear gradual change occurs.
3. determine which examples to keep and which to drop.
4. update models when significant changes is detected.

# Two kind of change in Physics data



(a)

(b)

Figure 1: (a) sudden change (b) gradual change.

## Basic idea

Basis idea :

1. It is motivated by the multiscale representation of functions [1]. We also represent data by the multiscale basis. The change in the raw data set is behavior as outlier in the coefficient of basis.
2. We transform sudden change or gradual change in raw data set to outlier in the coefficient set. To detect change, we only need to detect outlier in the coefficient set.

### Question

*How to transform the sudden change or gradual change to outlier?*

Using wavelet test function with certain order vanishing moment, which is describe in the book [1].

## Mathematical Basis

Let's use $f$ defined on $[0,1]$ to denote the wavelet test function. The support of $f$ is defined as

$$\operatorname{supp}(f) := \{x \in [0,1] : f(x) \neq 0\}$$

The wavelet test function $f$ has the following properties :

1. Vanishing moment [1] property of order $k$, that is

$$\int_0^1 f(x)x^j dx = 0, \; j = 0, 1, \ldots, k-1. \tag{1}$$

2. Local support, i.e., the support of $f$ is a subset of $[0,1]$.

We use $g$ to denote the target function we want to detect. The *coefficient set* is the set of the integration of target function $g$ and test function $f$, denoted as $\langle g, f\rangle$.

1. The property 1 is the key to distinct whether the changes happen or not. When the $f$ has order one vanishing moment,

   $|\langle g, f\rangle|$ close to 0 implies no change happens in the $\mathrm{supp}(f)$

   $|\langle g, f\rangle|$ far away from 0 implies some change happens in the $\mathrm{supp}(f)$

2. Property 2 can help us to locate where the change happens (not constant, not linear, etc) and the length of the support determines the accuracy the detection results.

# The choice of test function

For the sudden change, we use piecewise constant test function with order 1 vanishing moment.

$$f(x) = \begin{cases} 1, & 0 \leqslant x \leqslant 1/2 \\ -1, & 1/2 < x \leqslant 1 \end{cases} \tag{2}$$

For gradual change, we use piecewise linear test function with order 2 vanishing moment.

$$f(x) = \begin{cases} 1 - 4x, & x \in [0, \frac{1}{2}] \\ 4x - 3, & x \in (\frac{1}{2}, 1] \end{cases} \tag{3}$$

To obtain the local support test functions (different level test functions), we can shrink and shift the original test function (the details can be found in the book [1]).

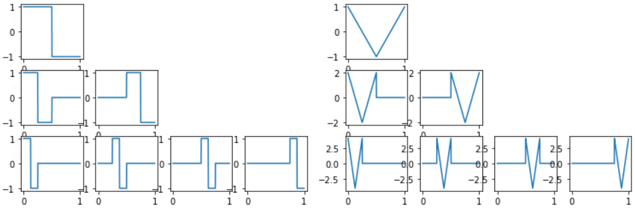


Figure 2: Local support test function with level 1, 2, 3.

## A simple example to explain the method

Let the target function $g$ is defined as

$$g(x) = \begin{cases} 0.8, & 0 < x < 0.6 \\ 0.2, & 0.6 < x < 1 \end{cases}$$

The test function, we choose is the thrid level of Haar wavelet (2)

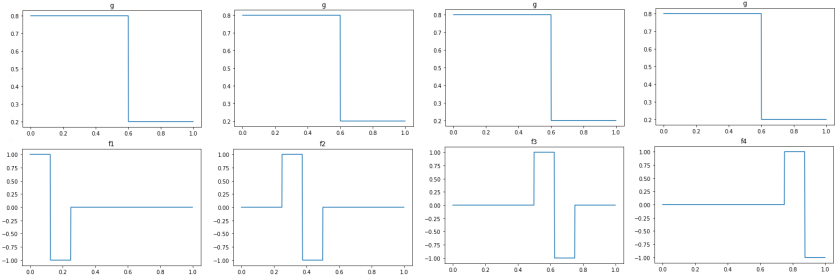


Figure 3: The target function and corresponding test function.

$$\langle g, f_1 \rangle = 0 \quad \langle g, f_2 \rangle = 0 \quad \langle g, f_3 \rangle = 0.06 \quad \langle g, f_4 \rangle = 0$$

$\langle g, f_3 \rangle = 0.06$ implies $g$ is not a constant in the support of $f_3$, i.e., [0.5, 0.75]. Thus, the result of our detection : $g$ is not a constant in the the support of $[0.5, 0.75]$, i.e., there are some change happens in $[0.5, 0.75]$.

## Change detect for Discrete Data

For continuous function $g(x)$, we check the integration of $g$ and local support wavelet test function $f$,

$$\int_0^1 g(x)f(x)dx \tag{4}$$

For discrete data, we use summation instead of integration. For discrete data,

$$[d_0, d_1, d_2, \ldots, d_{2^k-1}]$$

the wavelet test points

$$[x_0, x_1, x_2, \ldots, x_{2^k-1}]$$

are chosen as $x_k = \frac{i}{2^k} + \frac{0.5}{2^k}$, where $k$ is the level parameter. We check the summation,

$$a^k = \frac{1}{2^k} \sum_{i=0}^{2^k-1} d_i f(x_i) \tag{5}$$

We can check the amplitude of the summation to verify whether some changes happen in the data or not.

# Online Multiscale Change Detect

$$\underbrace{d_0, d_1}_{level1}$$

$$\overbrace{\underbrace{d_0, d_1}_{level1}\ \underbrace{d_2, d_3}_{level1}}^{level2}$$

$$\overbrace{\underbrace{d_0, d_1}_{level1}\ \underbrace{d_2, d_3}_{level1}}^{level2}\ \underbrace{d_4, d_5}_{level1}$$

$$\underbrace{\overbrace{\underbrace{d_0, d_1}_{level1}\ \underbrace{d_2, d_3}_{level1}}^{level2}\ \overbrace{\underbrace{d_4, d_5}_{level1}\ \underbrace{d_6, d_7}_{level1}}^{level2}}_{level3}$$

$$\dots$$

## Remark

1. *In each level, we check weighted the summation* (5). *If the summation is small, then we conclude no change happens in the data. If the summation is big (outlier), then we conclude some change (sudden change or linear change) happens in the data.*
2. *For the steaming data, we begin with the small level, and then increase the level when we obtain more data. The level parameter for discrete data is different with the level for continuous function.*
3. *If the change detected in a small level, then we can conclude the change happens in a small set, which means the detection results is more accuracy. So we start with accuracy level, and then detect for next level.*

## Data Description

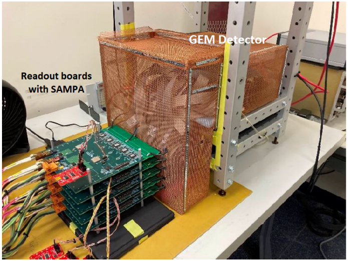


Figure 4: SAMPA front end cards and GEM detector [2]

1. The data we use the the channel 'moudle 1 strip 300 axis 1' from the data set 'hit_0_bbgem_304'. The data is obtain from SAMPA system and GEM detector in Jlab, and it is used to study cosmic ray.
2. The size of data is 30006912.

# Sudden change result

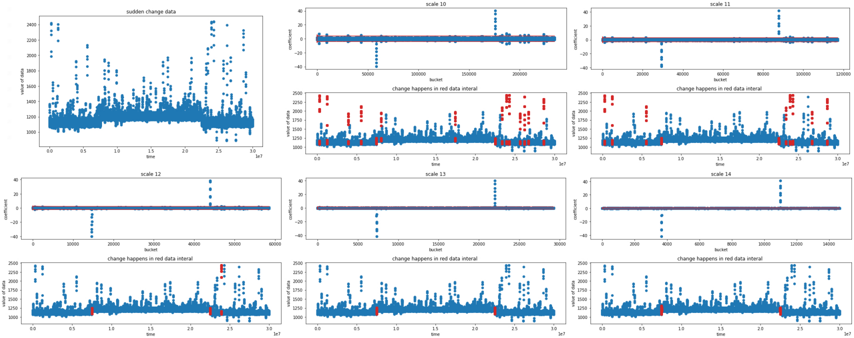


Figure 5: The results of online multiscale method for sudden change for the cosmic ray dataset.

## Sudden change result

1. the coefficient represent the summation of (5). For the sudden change, we choose the piecewise constant wavelet test function (2). The upper red and low red bound is determined by the mean and standard deviation of the coefficient.
2. If the coefficient lies out of the bound, we conclude change happens. The red point in the raw data set is the changeable data interval and the blue point the unchangeable data interval.
3. In the scale 10-11, most peaks has been detected in the algorithm. In scale 12-14, the sudden changes has been detected. So the accuracy for sudden is $2^{12} = 4096$ to $2^{14} = 16384$.

# Gradual change results

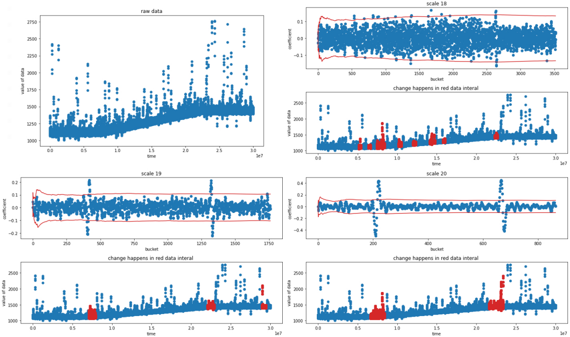


Figure 6: The results of online multiscale method for sudden change for the cosmic ray dataset

For the sudden change, we got a similar results. The gradual change is been detected in a higher level.

## Future Work

1. Once we have detected the change data, we need to calibration the changed data if the data is still useful.
2. If changed data which piece to keep and which piece to drop.

- Z. Chen, C. A. Micchelli, and Y. Xu, *Multiscale methods for Fredholm integral equations*, vol. 28, Cambridge University Press, 2015.

- E. Jastrzembski, D. Abbott, J. Gu, V. Gyurjyan, G. Heyes, B. Moffit, E. Pooser, C. Timmer, and A. Hellman, *Sampa based streaming readout data acquisition prototype*, arXiv preprint arXiv:2011.01345, (2020).

- I. Žliobaitė, M. Pechenizkiy, and J. Gama, *An overview of concept drift applications*, Big data analysis: new algorithms for a new society, (2016), pp. 91–114.