# 계절성 시계열 자료의 concept drift 탐지를 위한 새로운 창 전략

이도운 [1], 배수민 [2], 김강섭 [2], 안순홍 [2]
[12] 아시아나 IDT, AI 빅데이터연구소
leedw@asinaidt.com, baesm@asianaidt.com, kimks@asianaidt.com, ansh@asianaidt.com

# A novel window strategy for concept drift detection in seasonal time series

Do Woon Lee[1], Sumin Bae[2], Kangsub Kim[2], Soonhong An[2]
[12] ASIANA IDT, AI Big-data institute

**Abstract**

Concept drift detection on data stream is the major issue to maintain the performance of the machine learning model. Since the online stream is to be a function of time, the classical statistic methods are hard to apply. In particular case of seasonal time series, a novel window strategy with Fourier analysis however, gives a chance to adapt the classical methods on the series. We explore the KS-test for an adaptation of the periodic time series and show that this strategy handles a complicate time series as an ordinary tabular dataset. We verify that the detection with the strategy takes the second place in time delay and shows the best performance in false alarm rate and detection accuracy comparing to that of arbitrary window sizes.

## 1. Introduction

Different from the static environment in a laboratory, the real-world data have a dynamic and even time-variant characteristics. Since the concept of the real-world data changes, the necessity of concept drift detection arises. The traditional statistics to compare the distributions are hard to apply on especially time series because of their assumptions. For instance, KS-test is hard to apply on the time-series data because the distributions of datasets in a certain window scale are not independent. Because of this issue, previous research try to catch the drift in the manner of indirect approach[1].

We propose a novel window strategy with a perspective of Fourier analysis. In the case of time series with seasonality, the window size is set to be an integer multiple of their most probable period is the scale for the independent distributions.

## 2. Related work

Concept drift in time series is very different from other in various types of cases such as a classification problem in tabular data. This comes from the difference between the distribution of $p(X)$ and $p(X|y)$. In general, $p(X)$ and $p(y)$ are drawn from different distribution. But in case of time series, the domain of them are identical[2].

Drift detecting methods for time series with direct approach to data were proposed in previous researches[3][4]. However the industrial field has to detect the drift from a live stream, the preceding researches aimed to a re-collective way and the concern about the drift detection in time series is little.

## 3. Methodology

To define the size of window with Fourier decomposition, a pre-processing on time series is in necessary. The pre-processing for time series consist of following steps.

First, to avoid the irregular time interval, the data should be interpolated in piece-wise linear interpolation (or cubic Hermite spline interpolation). The optimal interval for the time series varies in each case. Next, the data is smoothed out with the Savitzky-Golay filter[5] to enhance the SNR (signal to noise ratio).

To find the exact period, decomposition is carried out for the sub-sampled intervals. Each sub-sampled interval gives a period, we calculate the $R^2$ score for every interval and set the most probable period with highest score. The window size is set to be the highest-scored period. The details are described in **Figure 1** as a form of pseudo code.

**Algorithm 1:** Window size computation

**initialization:** window size for `savgol_filter` $L_{SG}$
**input** : time series $(X, T) = \{x, t | (x, t) = (x_{col,i}, t_i)\}$
**output** : cycle of time series, $\omega$
**if** $len(T) \neq unique(T)$ **then**
  **for** $t_i \in T$ *and* $len(T = t_i) > 1$ **do**
    **for** $c \in columns$ **do**
      $X_{c,i} = \text{average}(X_c[T = t_i])$
    **end**
  **end**
**end**
$dt = \text{min}(T[1:] - T[:-1])$
**for** $c \in columns$ **do**
  $X_c = \text{interpid}(X_c, T)$
  $w_{SG} = 2L_{SG}/dt + 1$
  **if** $w_{SG} > 3$ **then**
    $X_c = \text{savgol\_filter}(X_c, w_{SG}, 3)$
  **end**
**end**
$N_{sub\_cycle} = \text{min}(\text{int}(\text{len}(T)/30), 50)$
$sub\_cycle = \text{random\_sample}(N = N_{sub\_cycle})$
**for** $sc \in sub\_cycle$ **do**
  $i_{end} = \text{int}(\text{len}(T) \times (1 - sc/2))$
  $freq = \text{FFT}(X_c[:i_{end}], T[:i_{end}])$
  $cycle = 1/freq_{\text{argmax}}$
  **if** $(cycle > 3dt)$ *and* $(cycle < (T_{max} - T_{min})/2)$ **then**
    $cycle\_cand \leftarrow cycle$
  **end**
**end**
**for** $cc \in cycle\_cand$ **do**
  $n = \text{int}(\text{floor}((T_{max} - T_{min})/cc))$
  **for** $i = 0$ *to* $n$ **do**
    **for** $j = i$ *to* $n$ **do**
      $score \leftarrow \text{R2\_score}(X_c[i : i + cc], X_c[j : j + cc])$
    **end**
  **end**
**end**
**return** $cycle[\text{argmax}(score)]$

**Figure 1. Window size algorithm**

## 4. Experiment

The experiments were conducted with 3 data sets. The details of the sets are shown in **Table 1**.

**Table 1** Time series data set description

| | *Period* | *Drift* |
|---|---|---|
| Climate data[6] | O | X |
| Sunspot data[7] | O | O |
| Synthetic data (1-10) | O | O |

Despite the concept drift is the key to maintain the performance of ML model, there are only few time series data sets with drift. And even for these sets, it is almost impossible to define the exact point of drift. This is the reason why synthetic data sets are included in the experiment.

Two real-world data set, Climate data and Sunspot data are seasonal time series from Max Planck Institute for Biogeochemistry and Royal Observatory of Belgium.

Despite of the *reality* of real-world data, these sets are lack of sign of drift, which can prevent an appropriate and quantitative analysis. On the other hand, synthetic data can give a chance to study the effect of drift and detection performance in this work.

The experiments were carried out in both real/synthetic cases, calculating D statistic in various window sizes. For the real-world data, this study explores Jena weather data set and the observation of the number of sunspot. In **Figure 2**, the climate data has no strong drift since it has a relatively short time range (2009 – 2017) in terms of weather. For this reason, there is no such turbulence in the statistics with the proper window size (365d, red solid line) comparing to that of an arbitrary one which is randomly selected (77 d, green dashed line).

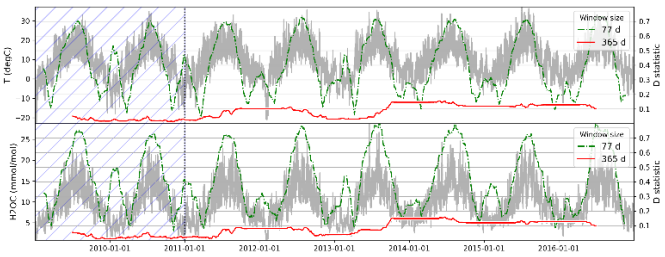


**Figure 2. Climate data from Jena, Germany. Gray solid line shows the Temperature (up) and H2OC (down). The hactch interval is the reference partition. The green dashed line shows D statistic with an arbitrary window size and red solid line is the one with proper size.**

On the other hand, the observation of sunspot number (SSN) over about 2 centuries shows subtle drift in between late 19th and early 20th century (**Figure 3**). As already well-known, the cycle of the solar magnetic activity (Schwabe cycle)[8] is about 11 years. However in **Figure 3**, there is an elevation of D statistic from 1880 to 1920. This change is certainly caused by a data drift because there is a long term cycle of ~179 years[9] in SSN and this larger one can't be seen in this time interval (1840-2020) with less than 2 period (~360 years). As shown in the **Figure 3**, suitable window size traces the changes in the data stream (red solid line) while randomly selected window size does not (green dashed line).

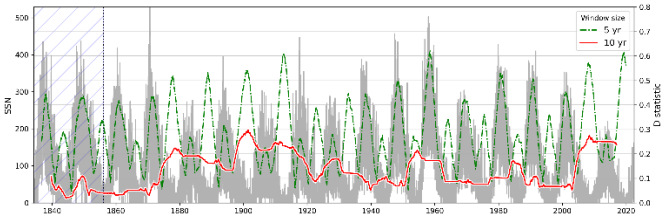


**Figure 3. The observation of sunspot number from 1818 – 2022. Gray solid line reprents the observation. Green dashed line and red solid line show the D statistic for 5yr, 11yr of window size.**

For the real-world data set, the data drift has a lack of quantitative evidence in the view of statistics though the drift has a solid physics on itself. Fortunately, synthetic data can point out the exact moment and strength when the drift occurred. 10 artificial periodic data set with the period of 365 days and different populations were used for the experiment. Window sizes were varied from 37 to 557 which are randomly selected. The data and D statistic as a function of time are shown in Figure 3, 1 to 5 panel. Data (X) is plotted as gray line and D statistics are shown as colored narrow lines. The hatched regions are the time range where the drift had happened. Black cross scatter is the time when the D statistic exceeds its 3σ. Depends on their window size, D statistic was

in a tranquil/undulating manner and these are summarized as a relative standard deviation ($C_v = \sigma/\mu$) of D. $C_v(D)$ is obviously small in the suitable (365 d) case whereas other cases show varying values.

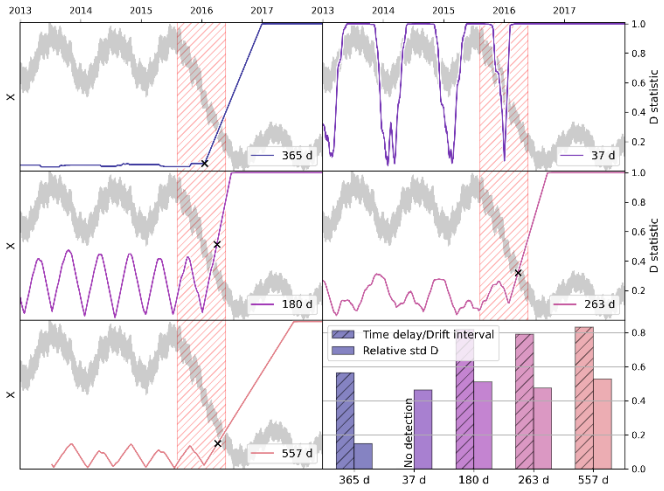


**Figure 4. D statistic as a function of time in various window sizes.**

**Figure 4** shows that the classical statistics with the window size of the exact period translates the time series as an ordinary data set. The window size of the period (365 d) is the case which has 0 no detection rate, the second shortest time delay (98 days) and the smallest $C_v(D)$ (0.17). These results indicate that an arbitrary size of window interprets the classical statistic as a meaningless one whereas proper size can trace the changes in data domain.

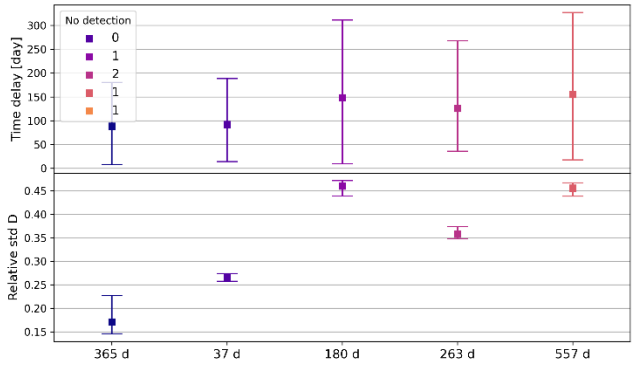


**Figure 5. Time delay and $C_v(D)$ in various window sizes.**

In summary, the arbitrary window size for the periodic time series can't trace down the change in data domain while the proper one can. This result shows that the classical approach with the window size of the period detect the drift if the target time series data has a seasonal characteristic.

## 5. Conclusion

This research explores the concept drift detection in time series data, both of real and synthetic sets. The key to adapt the classical statistics to periodic time series is the suitable window size. The window should be scaled with the period of the data, which lies in the frequency space.

In the experiments, the result shows that the arbitrary window size gives kind of false alarm whereas the proper size gives a clear sign of the drift with short time delay. Since the window size is set after the calculation of the window, the data can be treated as a tabular data. This result shows that the classical statistics with the proper window size successfully detect the drift in the periodic data.

There are further works to be explored in future. Our strategy can be applied in only periodic time series. The characteristic of time series only be ruled out when the data is in periodic manner. And there is an absent of a rule for threshold of D statistic. This depends on the user who build their model and the environment of the data.

## References

[1] Harries, M., & Horn, K. "Detecting concept drift in financial time series prediction using symbolic machine learning." AI-CONFERENCE-. World Scientific Publishing, 1995. P. 91-98

[2] R. C. Cavalcante, L. L. Minku and A. L. I. Oliveira, "FEDD: Feature extraction for explicit concept drift detection in time series", Proc. Int. Joint Conf. Neural Netw., pp. 740-747, 2016.

[3] L. Auret and C. Aldrich, "Change point detection in time series data with random forests," Control Engineering Practice, vol. 18, no. 8, pp. 990–1002, 2010.

[4] A. S. Block, M. B. Righi, S. G. Schlender, and D. A. Coronel, "Investigating dynamic conditional correlation between crude oil and fuels in non-linear framework: The financial and economic role of structural breaks," Energy Economics, vol. 49, pp. 23–32, 2015.

[5] Savitzky, A.; Golay, M.J.E "Smoothing and Differentiation of Data by Simplified Least Squares Procedures". Analytical Chemistry. 36 (8): 1627-39. 1964

[6] Max-Planck-Institut fur Biogeochemie, Jena Weather Station Saaleaue

[7] Sunspot data from the World Data Center SILSO, Royal Observatory of Belgium, Brussels

[8] Heinrich Schwabe, "Sonnenbeobachtungen im Jahre", Astronomische Nachrichten. 21: 233-236 from page 235. 1843

[9] Theodore J. Cohen, Paul R. Lintz, "Long erm peridocities in the sunspot cycle" Nature 250, 398-400. 1974