Dynamic Binning for the Unknown Transient Patterns Analysis in Astronomical Time Series

Thanapol Phungtua-eng
Department of Informatics
Shizuoka University
Shizuoka, Japan
thanapol@yy-lab.info

Yoshitaka Yamamoto
Department of Informatics
Shizuoka University
Shizuoka, Japan
yyamamoto@inf.shizuoka.ac.jp

Shigeyuki Sako
Institute of Astronomy
University of Tokyo
Tokyo, Japan
sako@ioa.s.u-tokyo.ac.jp

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671917

***Abstract*—In recent years, there arises a new opportunity for discovering transient phenomena such as supernovae, solar flares, and bursty events from detecting unknown transient patterns in astronomical time-series data. However, since these transient phenomena usually happen with unpredictable characteristics in shapes, sizes, and durations, scientists might lose some significant information due to the huge volume of astronomical data to be analyzed. Data sketching is useful to deal with such huge time-series data. A simple sketching technique is known as *binning* that captures the statistical summary of each bin of data points. In this paper, we attempt to provide a novel framework of data sketching for a statistical hypothesis testing and apply it for unknown transient pattern detection. The principal idea of statistical hypothesis testing lies in that two short-term and similar bins are mergeable into a long-term bin. By applying our proposed method, we suppress the unnecessary data while keeping the primary information without setting the bin size in advance. We evaluate our proposed method through experiments on the light curves in real-world data from telescopes with synthetic mixed-type transient patterns. Experimental results demonstrate that our proposed method outperforms several frameworks of transient pattern detection in astronomy.**

***Index Terms*—unknown transient pattern detection, dynamic binning, data stream, statistical hypothesis testing**

I. Introduction

The unknown transient patterns are phenomena that are significantly different and temporarily change from usual behavior [1], [2]. For unknown transient patterns observation, astronomers usually observe with an optical telescope. The optical telescope is used to measure the light intensity of stars in time series in the streaming manner. A sequence of light intensity measurements is called astronomical time series data. Astronomical time series data is a crucial asset to seek unknown transient patterns through aspects of computing. However, this data is usually complex data and involve the variety of external problems, such as atmospheric turbulence or measurement error from the observation hardware. The binning technique is simple and suitable for avoiding this problem.

The binning technique compresses the sub-interval of the astronomical time series data into the mean which is regarded as a statistical feature of the corresponding data series. However, astronomers cannot forecast the shape, height, or duration of transient patterns that can occur. Accordingly, it is infeasible to set a proper bin size for unpredictable characteristics of transient patterns in advance.

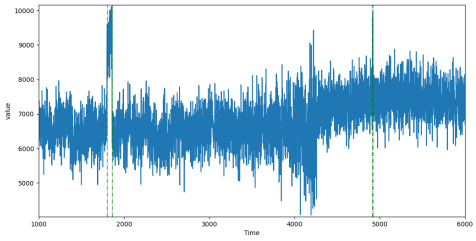


Fig. 1: Example of light curve data with two transient patterns

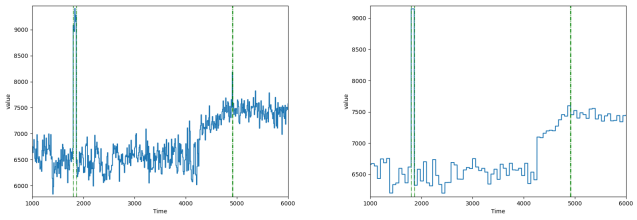
For example, Fig 1 shows the original astronomical time series data include both short and long duration, where the timestamps for the beginning of transient patterns are 4915 and 1800, respectively. In practice, we blindly fix bin size in advance under the constraint that the characteristic forms of transient patterns are unknown. Fig 2a demonstrates a sketch from the original data using the small bin size, which resulted in a failed attempt to remove noise. On the contrary, when the large bin size is used, the process compressed the short-duration unknown transient patterns with usual behaviors, which completely removed certain unknown transient patterns as shown Fig. 2b.

We address the key issue how to adequately sketch unknown transient patterns without defining the bin size from astronomers. This paper presents our novel technique to achiever an automatic bin size adjustment in a window called *dynamic binning* with the detection method.

II. Related Work

After previous investigations [4], we found that the most significant part of our method is dynamic binning. We propose to separate the dynamic binning from the detection process. Since we suppose dynamic binning that can capture correctly unknown transient patterns, it could improve the accuracy of the detection algorithm of [5].

Recently, Shevlyakov and Kan in [5] proposed transient pattern detection in Lunar Laser Ranging data. The input in their problem setting is a sequence of bins with a fixed bin size. Their bin was defined as the tuple (μ, s) where μ is



(a) Result of the sketch with bin size is equal to 10

(b) Result of the sketch with bin size is equal to 60

Fig. 2: Comparison of sketching with varying the bin size

the mean value of a sub-sequence light intensity measurement with interval T, s is the standard errors of a sub-sequence light intensity measurements with interval T, and T is the bin size that is defined by astronomers. We demonstrate the fixed bin size issue for detecting unknown transient patterns in Fig. 2. Our proposed method supports astronomers to adjust bin size properly and enables them to provide important features information in the bin.

There was another related work using feature extraction for sketching and applying to anomaly behavior detection. The feature extraction reduces unnecessary information and efficiently describes the important information. The independence is not obvious, the latest measurement of light intensity in time series is independent of the prior measurements. Therefore, it is practically impossible to extract features from astronomical time series data. Proposed methods to apply feature extraction to anomaly behavior detection are found in [3], [6].

III. Our proposed method

The aim of this section is to introduce our proposed method for automatic adaptive bin size. Firstly, we present the preliminaries of our research, as shown in Table I.

TABLE I: Table of notations.

Notations	Meanings
x_t	the measurement of light intensity at timestamp t
$X_{(1,t)}$	measurements of light intensity between from 1 to t $(x_1, x_2, \ldots, x_t)$
Bin	a tuple $(T, \mu, \sigma^2, \sigma^2 max)$ of $X_{(t,t+T)}$
W	the sequence of bins in a window $(Bin_1, Bin_2, \ldots, Bin_i)$
i	the index of Bin in W
μ_n	the mean of Bin_n
σ_n^2	the variance of Bin_n
$\sigma_{max_i}^2$	the maximum possible variance of Bin_i
α	the confidence level (initially set as 0.05)

The dynamic binning is an algorithm to adjust the proper bin size of each bin in window without defining a bin size. The merge algorithm is based on statistical hypothesis test two bins are mergeable.

Dynamic binning starts from a small bin size because it can describe the features of $X_{(1,t)}$ in detail. We put the x_t into Bin_{latest} when Bin_{latest} is full. Dynamic binning registers Bin_{latest} into W.

When Bin_{latest} is registering into W and W is not empty, dynamic binning considers merging Bin_{latest} and $Bin_{latest-1}$ in W into a new bin with a large size, if both bins are likely to be similar. We suppose the features of the two bins are not changed after merging two bins into the large bin. Dynamic binning is shown in Algorithm 1.

We test the means of two bins are equal by T-test at Line 1. T-test for equality of the means of two bins in the statistics is defined in Eq. 1. We suppose null hypothesis be $\mu_n = \mu_{n-1}$, and accept the null hypothesis is shown at Line 2. Where $T_{1-\frac{\alpha}{2}}$ is called critical value of t distribution with degrees of freedom.

$$\text{T-test} = \frac{\mu_i - \mu_{i-1}}{\sqrt{\sigma_i^2/n_i + \sigma_{i-1}^2/n_{i-1}}} \tag{1}$$

The means of the two bins are equal, but their variances can belong to different distributions. We apply F-test for equality of the maximum possible variances of two bins, which is defined in Eq. 2. The null hypothesis refers to maximum possible variances of two bins being equal, shown at Line 4. where $F_{1-\frac{\alpha}{2}}$ is called *critical value* of F distribution with degrees of freedom. If we accept the null hypothesis, we can merge two bins into new bin.

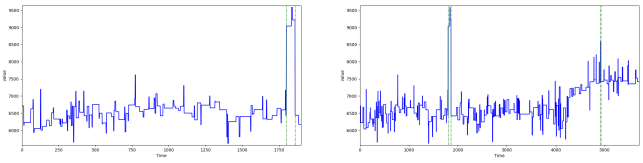
$$\text{F-test} = \sigma_{max_1}^2 / \sigma_{max_2}^2 \tag{2}$$

If two bins are not likely to be similar and the window is full, we search the bin which has the least T-test in the window then merge Bin_n and Bin_{n-1}. We compute this process to prevent the out-of-memory problem. The process is shown at Line 7-9.

Algorithm 1: *DynamicBin*

input: W: window, α: significance level

1 **Compute** T-test for equal means of Bin_l and Bin_{l-1}
2 **if** $|T_{test}| \leq T_{1-\frac{\alpha}{2}}$ **then**
3 **Compute** F-test for equal maximum possible variances of 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 **if** *W is full* **then**
8 Search the bin Bin_n whose T-test value is minimum in W
9 **Merge**(Bin_i, Bin_{i-1}).
10 **end**
11 **end**
12 **end**



(a) Result of sketching using dy- (b) Result of sketching using dynamic binning at timestamp 1900 namic binning at timestamp 5700

Fig. 3: Comparison of sketching using dynamic binning

IV. Empirical results

A. Lightcurve dataset

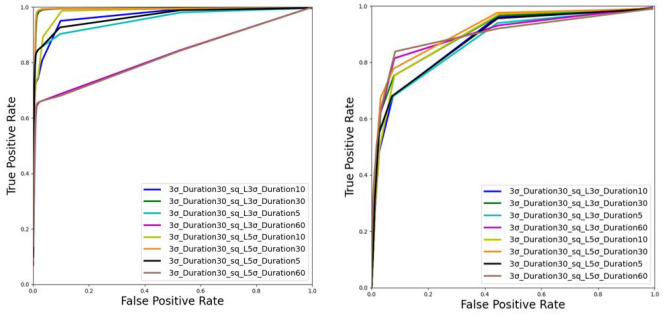
Astronomical time series data of this paper provides from a optical wide-field video observation system of Tomoe-gozen project. These 87000 datasets are generated from 476 files that are captured from optical telescope, by concatenating 476 files that each file consists of 2 unknown transient pattern whose height (amplitude) of a transient pattern is 3σ or 5σ from usual behavior and duration is defined from the instances ranged in [5, 10, 30, 60]. So ,each dataset contains 64 combination from previous size and duration

Fig. 1 shows an example of star file which contains two transient patterns. The duration and height of the first transient pattern is 60 and 5σ but the duration and height of the second transient pattern is 5 and 3σ.

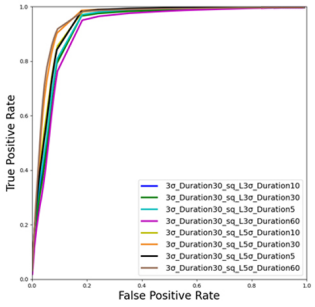
B. Empirical results

This subsection demonstrates an experimental evaluation of our proposed method. Firstly, we are interested in comparing the sketching result between dynamic binning and fixed bin size. Fig. 3 shows the results of sketching from dynamic binning with original data of Fig. 1. Our dynamic binning can capture the transient pattern at timestamp 1750. Then, when the second transient pattern appears, we can capture the second transient pattern at timestamp 4900. If we compare Fig. 3 and Fig. 2a, dynamic binning can filter out unnecessary information more than the small bin size. In the other hand, performance of noise filting of large bin size in Fig. 2b is better than dynamic binning but it cannot capture the second transient pattern. That causes false detections.

We also studied the accuracy rate of the transient pattern detection method using dynamic binning or fixed bin size. Fig. 4a and Fig. 4b, the result transient pattern detection method fixed bin of SK method [5]. When bin size and duration of transient patterns are equal but astronomers are unable to forecast the duration of the transient pattern happening in advance for defining bin size. Furthermore, when the bin size is not equal to any duration of transient patterns, the accuracy rate significantly drops. It is shown in Fig 4b. However, dynamic binning can use for any duration of transient patterns, as shown in Fig. 4c because dynamic binning can capture the transient patterns.



(a) ROC curves of SK method [5] with bin size equal to 30 (b) ROC curves of SK method [5] with bin size equal to 50



(c) ROC curves of SK method with dynamic binning

Fig. 4: Comparison ROC curve of fixed bin size and dynamic binning

V. Conclusion

In this paper, we propose dynamic binning to properly adjust bin size in the window and therefore can sketch unknown transient patterns with unpredicted duration. In further work, we will explore the analysis and sketch for transient patterns with more diverse and complex transient patterns. We hope our proposed method will be fruitful in discovering new phenomena in the real environment.

References

[1] Charu C. Aggarwal. *An Introduction to Outlier Analysis*. Springer International Publishing, Cham, 2017.

[2] Taegong Kim and Cheong Hee Park. Anomaly pattern detection for streaming data. *Expert Systems with Applications*, 149:113252, 2020.

[3] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 157–166, New York, NY, USA, 2005. Association for Computing Machinery.

[4] Thanapol Phungtua-eng, Yoshitaka Yamamoto, and Shigeyuki Sako. Transient patterns with unpredictable duration using chebyshev inequality and dynamic binning. In *2021 Ninth International Symposium on Computing and Networking Workshops (CANDARW)*, 2021.

[5] Georgy Shevlyakov and Margarita Kan. Stream data preprocessing: Outlier detection based on the chebyshev inequality with applications. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 402–407, 2020.

[6] Pengyuan Wang, Honggang Wang, Philip Hart, Xian Guo, and Kaveri Mahapatra. Application of chebyshev's inequality in online anomaly detection driven by streaming pmu data. In *2020 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, 2020.