A novel framework of non-parametric for adjusting the window size

Thanapiol Phungtua-eng[1,a)] Yoshitaka Yamamoto[1,b)]

Abstract: The data stream may contain irrelevant information due to various factors, such as the huge amount of data volume. The technique for removing irrelevant information is called *data binning*. The data binning is used to sequence the data stream into smaller bins and each of which captures statistical features in the corresponding sub-sequence. The obtained bins are collected into the window, meaning that the number of bins to be collected is determined by the window size. The window size is required to set in advance, whereas the sufficient value is varied with the target data stream to be captured. This paper proposes a novel framework for automatically adjusting the number of bins in the window with a non-parametric metric. We demonstrate our framework to detect unknown transient patterns with the astronomical data stream.

Keywords: Non-parametric data binning, Compression ratio, Astronomical data stream

1. Introduction

Astronomy is the exploration and discovery of unknown phenomena in the universe that bring more significant value for understanding the cosmos. The importance of this phenomenon is unexpected and occurs along with unpredictable behaviors, which is called *an unknown transient pattern*. The main characteristic of the unknown transient pattern in astronomy is a pattern that does not follow the prior observation of some representative periods [1], [5], [8], [10].

The data stream that refers to the behavior of celestial stars offers an opportunity to observe and study unexpected phenomena for great discovery in the Universe. However, analyzing live data stream by astronomers through manual inspection is improbable. Nowadays, the synergy in data science and astronomy is especially powerful for astronomical time series analysis and decision support. Data stream mining in time series is key to proper information extraction from large-scale data, which is widely used and of importance for applications in astronomy.

A mechanism in data stream mining applies unexpected behavior detection from prior evidence, called transient pattern detection [8].To analyze transient patterns in astronomy, we use light curve data that is the one of the more widely used research astronomical data for astronomers. The light curve data is the measurement from the photometric of a celestial object over a given period from optical telescopes at observation stations. Figure 1 illustrates the light curve data in a graph of light intensity from a certain star in time series.

The light curve data may contain *a confusion noise*, that becomes from the sensitivity of the optical telescopes generate a high Signal to Noise (S/N) due to awkward external factors [7]. To illustrate, the awkward external factors caused by the sky background, atmospheric noises, or measurement errors from hardware. The highlight of Figure 1 illustrates the unexpected behavior that suddenly increases and returns to normal. Based on the highlight, a question arises whether it is an ”unknown transient pattern” or a confusion noise. The confusion noise mainly causes astronomers to make wrong decisions. Generally, a simple approach to remove the confusion noise in astronomy is data binning. It is very important for time series summarization from original data and data stream mining.

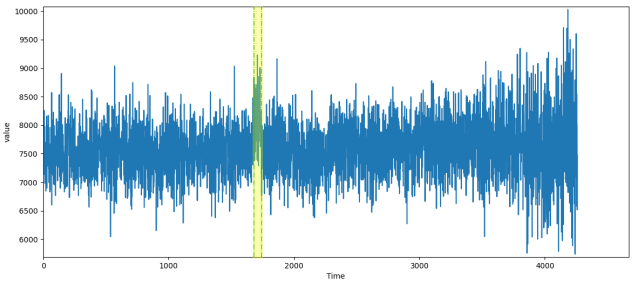
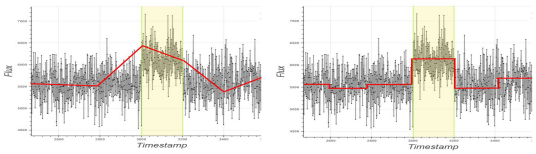


Fig. 1: An astronomical data stream from optical telescopes.

Data binning is a summary technique that reconstructs the original data into a new representation of fewer lengths, which are new representations of the original data [4], [12]. Afterward, the bins will be collected into *a window*, which refers to a memory resource of environment setting. The data binning requires a two-parameter model that consists of the maximum bin size and the maximum window size, which determine varying amounts of distortion from the shape or features of the original data. When astronomers define improper two-parameter, it loses of essential

[1] Department of Informatics, Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
[a)] thanapol@yy-lab.info
[b)] yyamamoto@inf.shizuoka.ac.jp



(a) Linear representation (b) Constant representation

Fig. 2: Comparison between linear and constant representation

original data features. Conversely, we decide to use the original data containing confusing noise corresponding to low-quality input. It cannot analyze transient patterns and makes wrong decisions. Defining Proper variables for a two-parameter is an issue to solve.

This research aims to solve problems regarding lossy compression and confusion noise removal with automatically the maximum bin size and the maximum window size defining. The core idea is non-parametric statistics that automatically specify proper variables based on statistical approaches. However, our preliminary study in that adjusting bin size was based on non-parametric statistics was presented in [11], it also called dynamic binning. We extend and improve the dynamic binning algorithm with the following contributions:

- We focus on improving non-parametric window size for dynamic binning. Additionally, we consider separating the transient pattern detection based on the Chebyshev inequality part from the compression part of the dynamic binning. Now, this non-parametric window size and dynamic binning apply to any detection.
- We propose the nonparametric statistics to adjust the input compression size that is proper for various unknown transient patterns and specific scenarios. The nonparametric statistics are used to support astronomers for time series analysis.

The remainder of this paper is organized as follows: Section 2 briefly reviews related work. We describe preliminaries defination and our proposed method to deal with the unknown transient patterns in Section 3. Then, Section 4 presents the results of our proposed method with the astronomical data stream from experimental evaluation. Finally, Section 5 gives conclusions and plans for future work.

2. Related Work

As a mentioned introduction, data binning is a summary technique that has applications for irrelevant information removal and reconstruction into bins. The typical technique for representing of bins in the time series consists of two types [6], shown in Figure 2.

The first type, linear representation for each bin, is a representation based on slope change that assumes a linear relationship of instances in the bin. The second type, constant representation for each bin, is a representation regarded as the important characteristic with constant that infer to the characteristics of the bin, which is generally a mean value of original data. Our specific domain is to discover sudden drastic changes and return to normal in the time series since a constant representation explains a sudden change better than a linear representation. Thus, our research focuses on the constant representation by the mean value for each bin.

Our compression's objective is to remove redundant and irrelevant data from confusion noise—conversely, compression affects distortion from the original data features by the representation length. We will measure how the fidelity results of data binning related to the original data. The simple measurement used to assess reconstruction quality is a compression ratio [3], [13]. The compression ratio is an essential measurement to define a proper representation length according to the variance loss caused by compression. Sulo, Berger-Wolf and Grossman proposed TWIN algorithm (Temporal Window in Networks) [13], which is a novel framework to optimal window size of data compression that can correspond to the original features of data in dynamic networks. However, the TWIN algorithm is based on fixed duration of window during the compression process. We did not develop our work based on fixed length of the representation

The landmark research of dynamic windowing technique (ADWIN: ADaptive WINdowing), is proposed by Bifet and Gavaldà [2]. This algorithm is based on an adaptive sliding window algorithm to detect change points. In our work, we aim to deal with the landmark window by automatically adjusting the variable size, which is an essential difference from ADWIN.

In conclusion, there is an important need for the non-parametric window size to apply with dynamic binning while also considering the balance between lossy compression and confusion noise removal by the compression ratio. In the next sections, we present the proposed method in details.

3. Our proposed method

In this section, we now introduce formal definitions after that we will present our proposed method.

3.1 Formal definitions

Definition 1: *astronomical data stream* is a sequence of measurement of light intensity in time series $X_{(1,T)} = (x_1, x_2, \ldots, x_t)$, where time interval is [1,T] and x_t denotes the latest measurement.

Definition 2: *Bin* is a representation of the statistical feature of sub-sequences of the astronomical data stream, which is divided into a small tuple by compression, denoted by Bin_i, where i is index-th of bin in the window. We formalize *Bin* with the following as $(n, \mu, \sigma^2, a, x_{max}, x_{min})$, where n is the number of instances. μ is the mean; σ^2 is the variance; a is the slope of bin calculated by linear least-squares regression; x_{min} is the minimum; and x_{max} is the maximum values in $X_{(m,m+n)}$, respectively. Note that we shall demonstrate the computing of all variables in the bin in Section 3.2.

Definition 3: *A window*, is a landmark window denoted by W and the maximum of window size is M. W collects the sequence of bins $\{Bin_1, Bin_2, \ldots, Bin_l\}$, that l is the latest bin such that $l \leq M$.

3.2 Data binning

Bin is a tuple to obtain a summary of features of the original data. The input of *Bin* is sub-sequences of the astronomical data

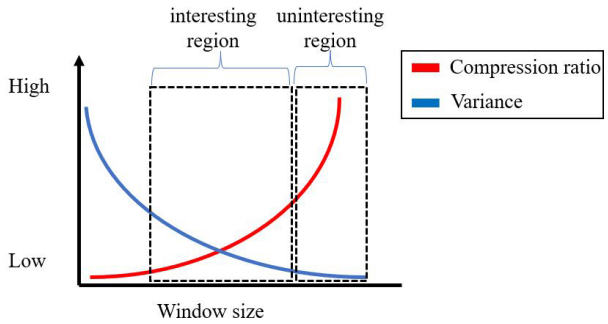


Fig. 3: Trade-off plot between compression and variance

stream ($X_{(m,m+n)}$). Let $X_{(m,m+n)} = \{x_m, x_{m+1}, \ldots, x_{m+n}\}$ where m is the timestamp that is the starting point of the sub-sequence, and $m + n$ is the endpoint of the sub-sequence. The structure of the bin consists of six metrics.

(1) Number of instances (n): It is the sub-sequences length of the astronomical data stream that is compressed into a bin.

(2) Means (μ): It is a representative of a list of numbers:

$$\mu = \frac{\sum_{i=m}^{m+n} x_i}{n} \tag{1}$$

(3) Variance (σ^2): It measures how far instances in the bin are spread out:

$$\sigma^2 = \frac{\sum(x_i - \mu)}{n-1} \tag{2}$$

(4) Slope of bin (a) : It measures how steep the line is:

$$a = \frac{N(\sum x \cdot t) - (\sum x)(\sum t)}{n(\sum x^2) - (\sum x)^2} \tag{3}$$

(5) Minimum value in the sub-sequence (x_{min}) : It corresponds to a minimum value in the sub-sequences astronomical data stream is compressed into a bin.

(6) Maximum value in the sub-sequence (x_{max}): Conversely, it corresponds to a maximum value in the sub-sequences astronomical data stream

Example Let $T = 5$ and $X_{(1,5)} = \{200.00, 300.00, 400.00, 500.00, 600.00\}$, thus, the *Bin* correspond to $X_{(1,5)}$, represented in tuple {5, 400.00, 25000.00, 100, 200.00, 600.00}.

3.3 Landmark window

The window is part of contributions to manage the sequence of bins into memory. The landmark window is a model that contains all of the bins from the starting timestamp to the latest timestamp, and bins in this window are equally important [9]. When the window is full, and Bin_l is entering to window, the dynamic binning consider merging two bins that are neighborhood and similar. Notably, the dynamic binning is described in Section 3.5.

The limitation of the landmark window is that it is difficult to define the proper size by users. We propose a mechanism to adjust the proper size, which is based on the compression ratio and variance.

3.4 Trade-off compresion ratio and variance

A simple way to solve the issue of proper window size is to increase it until it reaches the expected result. However, a common question is whether a specific window size determining appropriate for this scenario. For example in the worst-case scenario, a low variance also means low information. The reasonable metric is the compression ratio, which measures the relative reduction in the size of data representation. Thus, we consider the compression ratio as an essential measurement for the judgment of window size because the variance and compression ratio do have opposite behaviors. Sulo, Berger-Wolf, and Grossman [13] explain the trade-off between compression ratio and variance as shown in Figure 3.

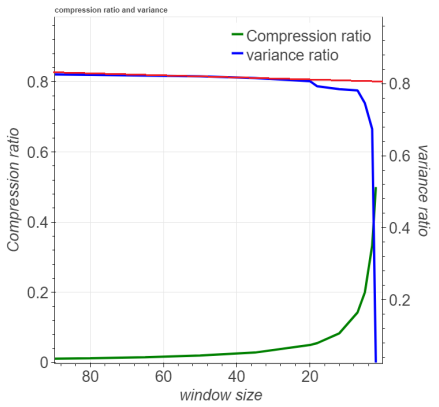


Fig. 4: Example of threshold line in trade-off graph.

Small variance and large compression ratio correspond to the small window size, which has too few bins for the representation of original data without lossy information, called uninteresting region in the Figure 3. Indeed, we should balance the window size in the interested region, corresponding to high information with a proper window size. The compression ratio is the ratio between the length of the original data and the window size. The compression ratio of window (R) is defined as:

$$\mathrm{R}_w = \frac{T}{c_w} \tag{4}$$

T denotes the length of $X_{(1,T)}$, and M_w denotes the window size that is defined by users. The main contribution of finding the proper window size is to be used for transient pattern detection. The input data, which has high information and a small number of bins in the window, makes it easy to make decisions whether it is a transient pattern.

For example, let $X_{(1,100)}$ is input, and we compress them into two bins in the window. The value of the compression ratio is 50, and the variance may be low because the population for variances computing is equal to two instances. Thus, the result of compression may not be insufficient to analyze the transient pattern. This technique finds an appropriate window size of the base shape of the original data, reducing variance and irrelevant data, and without lossy compression.

3.5 Our proposed method

Our proposed method automatically specifies proper representation size (bin size and window size) for any transient pattern detection application. The requirement of the real environment consists of two steps. The first step is adjusting bin size dy dynamic binning

3.5.1 Dynamic binning

The dynamic binning is suitable for compressing the light curve data as it continuously measures and records the light intensity into bins. In this situation, we do not know the length of the original data since they still continually measure light intensity. Thus, the core strategy is to merge two small bins that are similar into the large bin by statistical hypothesis testing. The algorithm to adjust the bin size was published in [11], which describes the pseudocode of our proposed online method.

3.5.2 Adjusting window size

Defining the length of output from the compression without lossy compression is the main problem of our research. The goal of this paper is to focus on the result of compression that infers the original data features, and the output must correspond to high information. We select the window size from the point where variance is suddenly dropping from the threshold line. For example, we demonstrate the trade-off graph and threshold line, where the window size equals 20 is the starting point of suddenly dropping variance, as shown in Figure 4

In the experiment of our proposed method, we performed the evaluation of the data binning with window size in the interesting region by capturing transient patterns that are a variety of scenarios in Section 4, which studies the influence of bin size and window size.

4. Experiments

In this section, we experimentally evaluate the performance of the our proposed method. Our proposed method was implemented in Python 3.9.2. The evaluations of this study consider three different objectives:

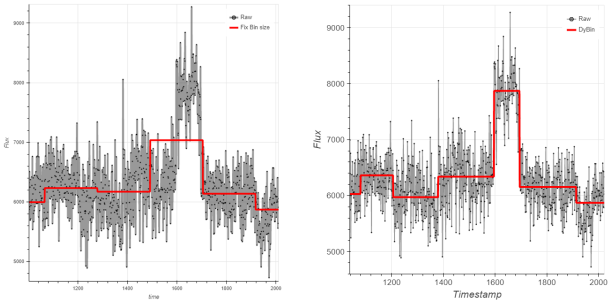
- Our domain is a real-world problem that applies to the optical telescope. We evaluated our proposed method using light curve data generated from a real environment.
- We aim to influence the window size and evaluate their automatic adjustment to compress various unknown transient patterns with compression ratio by dynamic binning.
- Our proposed method also removes redundant and irrelevant data from confusion noise with lossless compression to support the decisions of astronomers for unknown transient pattern detection.

4.1 Light curve data

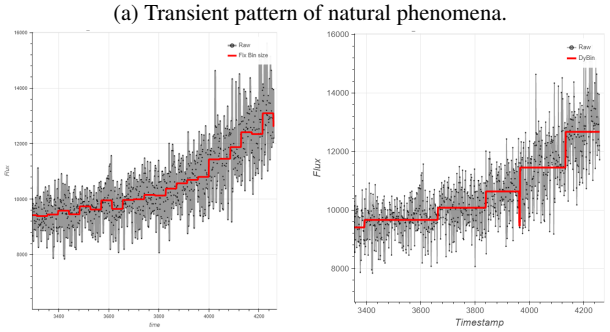
The light curve data were generated using an optical wide-field video observation system composed of a mosaic CMOS camera on the Kiso Schmidt telescope [*1]. Notably, we removed the background noise using the PCA methods in all light curve files.

We refer to this set of light curves as the real transient patterns of natural phenomena in a real environment and were captured using the optical telescope. The dataset consists of files containing real transient patterns or expected behavior. To address our objective, we focus on compression results using various bin sizes and window sizes for transient patterns capturing and confusion noise removal. We evaluate with the following four scenarios that we found:

[*1] The dataset provided is from the Tomoe-gozen project. For more detail, visit https://tomoe.mtk.ioa.s.u-tokyo.ac.jp/



(a) Transient pattern of natural phenomena.



(b) Gradual changing

Fig. 5: Comparison between unknown transient pattern and gradual changing.

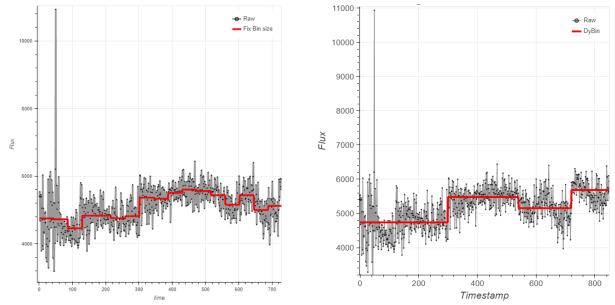
- Normal behavior: This period corresponds to near-constant stable data, and it can be noisy.
- Transient pattern (sudden change): This can happen suddenly/abruptly in a period and then return to normal behavior.
- Outlier point: This is an outlier point that may become from external factors. It may happen by a few instances on the data stream.
- Gradual changing: This period has slowly changed over time, and it may become new normal behavior. It is one kind of pattern in astronomy. However, this is not our objective.

4.2 Fixed bin size vs. dynamic bin size in compression

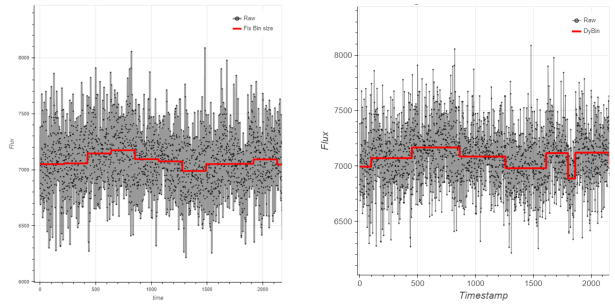
We carried out an experiment on proper bin size in compression for various scenarios with lossless compression from the original feature. Our experimental goal of proper bin size in compression is to justify the bin size automatically. We used the four scenarios of light curve that are distinctly different features for comparison: normal behavior, unknown transient pattern, outlier point, gradual change. To perform a fair comparison between fixed bin size and dynamic bin size, we fix the window size that is in the interesting region (Figures 3) and plotted the result of compression by binning, as shown in Figures 5 and 6. The black and red lines represent the original data and results from compression, respectively.

The dynamic bin size can adjust the proper bin size for a transient pattern capture. The small bin size infers a sudden changing feature, and the large bin size infers the period in which the original data have a stable period, as shown in Figure 5a.

We found a minor issue in the gradual change scenario with a small bin size during a slow change in Figure 5b. Because we initialize with the smallest bin size and then cannot merge

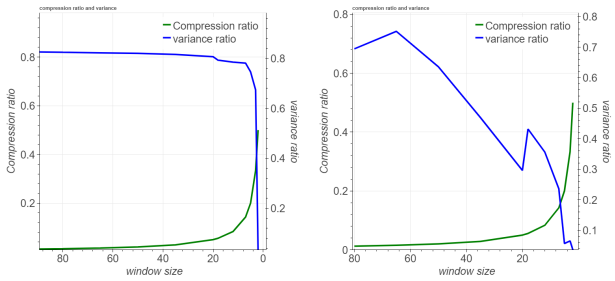


(a) Outlier point



(b) Normal behavior

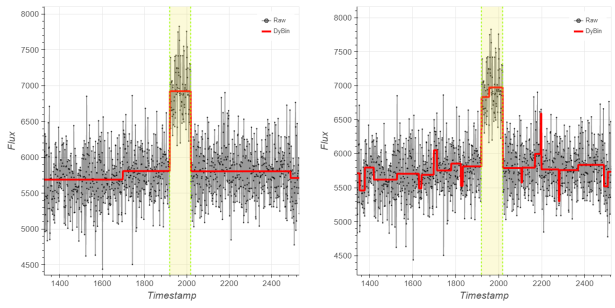
Fig. 6: Comparison between outlier point and normal behavior



(a) Transient pattern with dynamic bin size.

(b) Transient pattern with dynamic bin size.

Fig. 7: Comparison trade-off variance and compression ratio graph between the fixed bin and dynamic bin size.

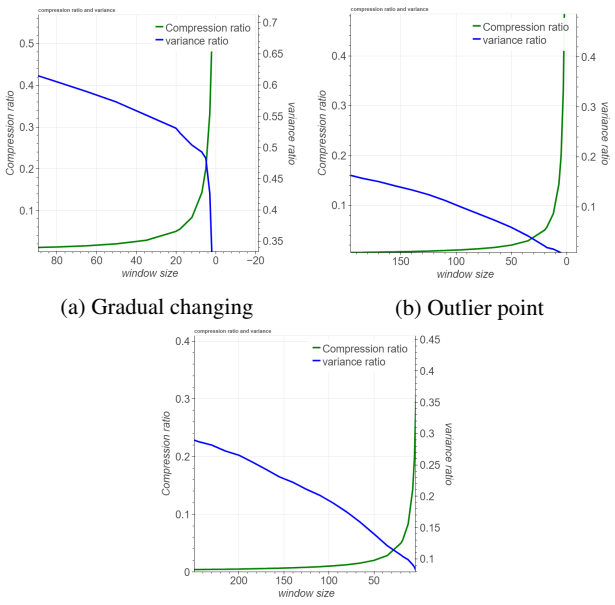


(a) window size = 10

(b) window size = 70

Fig. 8: Comparison of dynamic binning with variety window size.

with neighbor bins by statical hypothesis testing. Additionally, the window size in the interesting region demonstrates to support dynamic binning, which overall results are similar. However, the gradual change scenario is not our objective, and the dynamic binning represents a successful attempt to capture unknown transient patterns.



(a) Gradual changing

(b) Outlier point

(c) Normal behavior

Fig. 9: Trade-off variance and compression ratio graph of other scenarios.

4.3 Influence of the window size on compression results

In the preceding subsection, the results in Figures 5 and 6 indicate that the dynamic bin size, which automatically adjusts the bin size via dynamic binning, is appropriate for sketching unknown transient patterns without lossy compression. This subsection describes the experiments related to the window size's influence on the interesting region of the trade-off between the compression ratio and variance.

Figure 7 shows the trade-off graph comparing fixed and dynamic bin sizes. The above results indicate that dynamic binning appropriately compresses the transient pattern, and we compare the result of trade-off variance and compression ratio, as shown in Figures 7a and 7b. We can observe that the variance fixed bin size slowly decreases. The reason is that the bin size increased when the window size decreased. The mean of the bin that contains the peak signal of the transient pattern is reduced by increasing the bin size. In addition, the transient pattern disappears by compression into the larger bin.

Conversely, we found a variance ratio that suddenly increases in Figure 7b then drops again. This is because the fixed bin size can capture the transient pattern. However, it is difficult to specify window size, which depends on the bin size, in order to capture the transient pattern compared to the non-parametric window size with dynamic binning.

When we compress with the other scenarios, compression results with dynamic binning demonstrate adjusting window sizes, as shown in Figure 9, which can be adapted by defining the threshold of variance reduction. The variance in the trade-off results slowly decreases when we reduce the window size, which reduces the noise from the original data features. The compression results by dynamic binning are unsatisfactory. However, the unknown transient pattern is more important in our objective than in other scenarios.

In summary, nonparametric window sizes with unknown tran-

sient patterns can freely define window sizes without cornering significant differences results, as shown in Figure 8 (the highlight is the period of unknown transient pattern).

5. Conclusions

This paper introduces a non-parametric method for adjusting the window size with dynamic binning. Our empirical experiments have shown that compression with the automatic adjusting bin size and window size accomplished without lossy compression from the original data features and removing confusion noise. The trade-off between compression ratio and variance demonstrates that when the window size is not significantly small, the compression results may not be distorted from the features of the original data. The window size is essential for analyzing and exploring transient patterns. Although we found a gradual change scenario, it may be a phenomenon in astronomy that is out of our domain. We need to address this issue in future work.

Furthermore, we demonstrated the evaluation of real transient patterns of natural phenomena with detection application. We will explore the data binning with other scenarios and extend our proposed method to detection based on clustering. Moreover, we also expect further improvements to our proposed method to discover new phenomena in a real environment.

Acknowledgments

We are grateful to Shigeyuki Sako (University of Tokyo) for providing the dataset and guidance that made this research successful.

References

[1] Aggarwal, C. C.: *An Introduction to Outlier Analysis*, Springer International Publishing (2017).

[2] Bifet, A. and Gavaldà, R.: Learning from Time-Changing Data with Adaptive Windowing, *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA*, SIAM, pp. 443–448 (online), DOI: 10.1137/1.9781611972771.42 (2007).

[3] Chiarot, G. and Silvestri, C.: Time series compression: a survey (2021).

[4] Cormode, G. and Yi, K.: *Small Summaries for Big Data*, Cambridge University Press (2020).

[5] Gama, J. a., Žliobaitundefined, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A.: A Survey on Concept Drift Adaptation, *ACM Comput. Surv.*, Vol. 46, No. 4 (online), DOI: 10.1145/2523813 (2014).

[6] Goldstein, R., Glueck, M. and Khan, A.: Real-time compression of time series building performance data, *Proceedings of Building Simulation*, pp. 14–16 (2011).

[7] Helou, G. and Beichman, C. A.: The confusion limits to the sensitivity of submillimeter telescopes, *Liege International Astrophysical Colloquia* (Kaldeich, B., ed.), Liege International Astrophysical Colloquia, Vol. 29, pp. 117–123 (1990).

[8] Kim, T. and Park, C. H.: Anomaly pattern detection for streaming data, *Expert Systems with Applications*, Vol. 149, p. 113252 (online), DOI: https://doi.org/10.1016/j.eswa.2020.113252 (2020).

[9] Mansalis, S., Ntoutsi, E., Pelekis, N. and Theodoridis, Y.: An evaluation of data stream clustering algorithms, *Stat. Anal. Data Min.*, Vol. 11, pp. 167–187 (2018).

[10] Martínez-Galarza, J. R., Bianco, F. B., Crake, D., Tirumala, K., Mahabal, A. A., Graham, M. J. and Giles, D.: A method for finding anomalous astronomical light curves and their analogues, *Monthly Notices of the Royal Astronomical Society*, Vol. 508, No. 4, p. 5734–5756 (online), DOI: 10.1093/mnras/stab2588 (2021).

[11] Phungtua-Eng, T., Yamamoto, Y. and Sako, S.: Detection for Transient Patterns with Unpredictable Duration using Chebyshev Inequality and Dynamic Binning, *2021 Ninth International Symposium on Computing and Networking Workshops (CANDARW)*, pp. 454–458 (online), DOI: 10.1109/CANDARW53999.2021.00084 (2021).

[12] Sayood, K.: 1 - Introduction, *Introduction to Data Compression (Third Edition)* (Sayood, K., ed.), The Morgan Kaufmann Series in Multimedia Information and Systems, Morgan Kaufmann, Burlington, third edition edition, pp. 1–11 (online), DOI: https://doi.org/10.1016/B978-012620862-7/50001-8 (2006).

[13] Sulo, R., Berger-Wolf, T. and Grossman, R.: Meaningful Selection of Temporal Resolution for Dynamic Networks, *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, New York, NY, USA, Association for Computing Machinery, p. 127–136 (online), DOI: 10.1145/1830252.1830269 (2010).