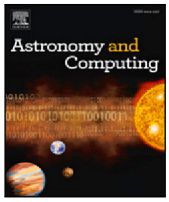
Astronomy and Computing 45 (2023) 100765

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom



Full length article

Effective detection of variable celestial objects using machine learning-based periodic analysis

N. Chihara [a,*], T. Takata [b,c], Y. Fujiwara [d], K. Noda [e], K. Toyoda [e], K. Higuchi [e], M. Onizuka [a]

[a] Osaka University, Japan
[b] The Graduate University for Advanced Studies, SOKENDAI, Japan
[c] National Institutes of Natural Sciences National Astronomical Observatory of Japan, Japan
[d] NTT Communication Science Laboratories, Japan
[e] TDAI Lab Co., Ltd., Japan

ARTICLE INFO

Keywords:
Variable celestial objects
Periodic analysis
Sparse modeling
Anomaly detection

ABSTRACT

This paper tackles the problem of effectively detecting variable celestial objects whose brightness periodically changes over time. This problem is crucial in studying the evolution and structure of the universe and elucidating physical phenomena. The method by Sesar et al. is one of the popular approaches used in detecting variable celestial objects that uses statistical data of celestial time series, such as intrinsic variability σ and χ^2, etc. However, since statistical data is an aggregation of celestial time series, the previous approaches do not take advantage of the periodicity, which is the inherent characteristic of variable celestial objects; it fails to find variable celestial objects effectively. To solve such a problem, we propose an approach to detecting variable celestial objects using periodic analysis. Our approach uses sparse modeling as periodic analysis since celestial time series is typically sparse and sparse modeling can effectively obtain periodicities of the celestial objects from sparse time series. By exploiting the periodicities of the celestial objects as features, we perform binary classification to estimate whether a celestial object is a variable celestial object. To show the effectiveness of our approach, we evaluated our approach using Hyper SuprimeCam (HSC) PDR2 dataset, and we confirmed that AUC of our approach is 0.939 while AUC of the previous approach is 0.750; our approach can more effectively detect variable celestial objects.

1. Introduction

Variable celestial objects refer to objects in the sky whose brightness changes over time. They are revealed to account for approximately several percent of the total celestial objects detected in the sky surveys (Sesar et al., 2009; Bhatti et al., 2010), although the fraction is dependent on the survey depth and observation wavelength (filters). Their representative examples are flare stars, novae, supernovae, periodically variable stars such as Cepheids, RR Lyrae stars, and so on, and also active galactic nuclei(AGN), etc. The way of variation in their brightness on variable objects is a vital and significant phenomenon in revealing the structure and evolution of the universe. Specifically, investigating variable celestial objects allows us to probe cataclysmic and catastrophic events and mass accretion by massive black holes (Gosnell et al., 2022; Yan-Ke et al., 2022; Eyer and Blake, 2005; Ofek et al., 2020). In addition, the period of periodically variable objects can provide some important information, such as distance to the objects (Sesar et al., 2009; Braga et al., 2021; Liu et al., 2022) for probing the structure of our galaxy, for identifying the faint and small nearby galaxies (Vivas et al., 2022) and radii and mass of eclipsing M dwarf systems followed by spectroscopic radial velocity measurements (Becker et al., 2011). So, many astronomical researchers search for new variable celestial objects and investigate their characteristics. On the other hand, the importance of effective detections of variable objects is increasing, as many wide field imaging surveys with large telescopes did and will produce huge datasets including numerous variable objects, and fast and accurate selections of variable objects are essential for making dramatic progress in the knowledge of the active universe.

A statistics-based detection is a major approach for detecting variable celestial objects. For example, the method by Sesar et al. (2007) identifies variable celestial objects using statistics, such as intrinsic variability σ, χ^2, etc. However, since such low-order statistics are aggregations of astronomical time series, this method fails to take

* Corresponding author.
E-mail addresses: naoki88@sanken.osaka-u.ac.jp (N. Chihara), tadafumi.takata@nao.ac.jp (T. Takata), yasuhiro.fujiwara@ntt.com (Y. Fujiwara), koki.noda@tdailab.com (K. Noda), keisuke.toyoda@tdailab.com (K. Toyoda), kaito.higuchi@tdailab.com (K. Higuchi), onizuka@ist.osaka-u.ac.jp (M. Onizuka).

https://doi.org/10.1016/j.ascom.2023.100765
Received 19 July 2023; Accepted 19 October 2023
Available online 3 November 2023

Symbols and definitions

Periodic analysis

n	Length of astronomical time series
m	Number of component frequencies
N	Number of astronomical time series
f_{nyq}	Nyquist frequency
X	Explanatory variables $X \in R^{n \times 2m}$
$\boldsymbol{y}$	Target variables $\boldsymbol{y} \in R^n$
$\mathcal{Y}$	Astronomical time series dataset $\mathcal{Y} = \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_N}\}$
$\boldsymbol{w}$	Regression coefficient $\boldsymbol{w} \in R^{2m}$
$\boldsymbol{\nu}$	Frequency candidates $\boldsymbol{\nu} \in R^m$

Group Lasso

J	Number of group
g_j	Number of variables of jth group $(1 \le j \le J)$
$\boldsymbol{G_j}$	Indices of jth group $\boldsymbol{G_j} \in R^{g_j}$
$\mathcal{G}$	Group set $\mathcal{G} = \{\boldsymbol{G_1}, \ldots, \boldsymbol{G_J}\}$

Binary classification

$\mathcal{F}_p$	Periodic feature $\mathcal{F}_d \in R^{N \times (m+1)}$
$\mathcal{F}_d$	Domain feature $\mathcal{F}_d \in R^{N \times 4}$
$\mathcal{F}$	Feature list $\mathcal{F} \in R^{N \times (m+5)}$
$\mathcal{L}$	List of variable celestial objects

advantage of the periodic pattern, which is the inherent nature of variable celestial objects. In this paper, we utilize the periodic pattern to identify periodically variable objects. A typical method for extracting the periodic pattern is the Fourier transform. This method needs to impute missing values in the astronomical time series for each celestial object to extract the periodic pattern. However, the ratio of missing values in astronomical time series (e.g. HSC PDR2) is quite high (Aihara et al., 2019) (See Fig. 1). Therefore, this method does not work well for the astronomical time series. We tackle this problem by leveraging a sparse modeling technique, which is a machine learning-based approach robust against missing values. This technique can train robust models by extracting relevant variables from sparse time series. In addition, we design input features of binary classification to accurately estimate whether a celestial object is a variable celestial object by investigating the nature of variable celestial objects. In order to process large-scale astronomical time series efficiently, we employ distributed processing for the sparse modeling technique. In short, we summarize the key contributions of this work as follows:

Effective: The proposed method effectively detects variable celestial objects by extracting their representative periodic pattern, which is their inherent characteristic. It can automatically extract representative periodic patterns by leveraging sparse modeling.

Robustness: Astronomical time series have a large number of missing values, and they are so sparse due to the observation environment. Our method for analyzing such series exhibits robustness against sparse astronomical time series by leveraging sparse modeling to mitigate overfitting.

Scalability: The proposed method independently processes each celestial object. Therefore, it is capable of high-speed processing for millions of celestial objects by employing distributed processing.

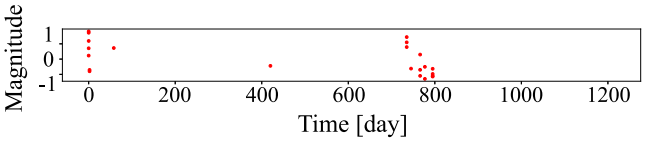


Fig. 1. An example of an astronomical time series such that the magnitudes of variable celestial objects periodically changes over time. The x-axis denotes time, while the y-axis denotes standardized magnitude. Astronomical time series in the HSC PDR2 dataset are extremely sparse.

2. Preliminary

We describe sparse modeling techniques. They can extract relevant variables for representing high-dimensional data like astronomical time series, eliminating redundant variables that potentially cause overfitting. In particular, we introduce Lasso regression and its variant, Group Lasso regression as follows.

Lasso (Tibshirani, 1996): it is a linear regression model that has become a popular feature selection and shrinkage estimation method. The loss function of the Lasso estimator E_L is defined as

$$E_L = \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \tag{1}$$

In this equation, $X \in R^{n \times m}$ is explanatory variables, where n is the length of astronomical time series and m is the number of component frequencies, $\boldsymbol{y} \in R^n$ is target variables, $\boldsymbol{w} \in R^m$ is regression coefficient, and $\|\boldsymbol{w}\|_1 \equiv \sum_i |w_i|$. According to the penalty parameter, λ, several coefficients of $\mathbf{w}$ are set exactly to zero; coefficients becoming zero have no involvement in the model. Moreover, continuous shrinkage can improve the model accuracy due to the bias–variance trade-off (Li et al., 2011). Such a process of removing irrelevant or redundant variables is commonly referred to as feature selection. The success of feature selection depends on ℓ_1 penalty, which is the second term in (1) feature selection reduces the model's complexity in the astronomical time series; it mitigates overfitting.

Group Lasso (Yuan and Lin, 2006): The Group Lasso is an extension of the Lasso regression, which performs the feature selection on predefined groups of variables in linear regression models. The loss function of the Group Lasso estimator E_{GL} is defined as

$$E_{GL} = \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{w}\|_2^2 + \lambda \sum_{\boldsymbol{G}_j \in \mathcal{G}} \|\boldsymbol{w}_{\boldsymbol{G}_j}\|_2 \tag{2}$$

In this equation, $\mathcal{G} = \{\boldsymbol{G_1}, \ldots, \boldsymbol{G_J}\}$ is the group set representing how m variables are divided into J groups where $\boldsymbol{G_j}$ is the index set belonging to the jth group of variables. The penalty in the second term of (2) can be viewed as an intermediate between the ℓ_1 and ℓ_2 penalty. Therefore, it performs feature selection at the group level and is invariant under group-wise orthogonal transformations like ridge regression (Hoerl and Kennard, 1970). Note that when $\lambda = 0$, the above equation is identical to linear regression.

3. Proposed method

We describe the technical challenges and our approach to detecting variable celestial objects.

3.1. Technical challenges and problem formulation

There are two technical challenges in detecting variable celestial objects using astronomical time series. The first challenge is that the astronomical time series have an extremely large number of missing values (See Fig. 1 as an example). This prevents applying standard time series analysis, such as Fourier transformation. This is because the standard time series analysis does not work with any missing values, and we need to use data imputation for them. However, data

imputation does not work well because the amount of missing values in astronomical time series is overwhelming. To overcome this problem, we utilize sparse modeling techniques that can train robust models by extracting relevant variables from sparse time series. In detail, we first extract representative periodic patterns (frequencies) from astronomical time series datasets by sparse modeling and then detect variable celestial objects using the extracted frequencies. In order to detect variable celestial objects, we utilize the representative frequencies and the quality (score) of the frequency extraction. The second challenge is how we can design input features of binary classification for effectively detecting variable celestial objects. In addition to utilizing the representative frequencies, we carefully design additional input features by investigating the nature of celestial objects. For example, we use the celestial object's magnitude as the additional feature. We easily obtain periodic patterns from bright celestial objects, on the contrary, it is difficult to obtain them from dark ones. So, if we take advantage of the magnitude of celestial objects as one of the input features, the accuracy of binary classification expects to be higher.

Based on the two challenges mentioned above, we formulate the problem of detecting variable celestial objects using time series datasets as follows:

$$\mathcal{L} = F(\mathcal{Y}, m, f_{\mathrm{nyq}}, \mathcal{F}_d) \tag{3}$$

where F is a function that detects variable celestial objects $\mathcal{L}$ from the astronomical time series dataset $\mathcal{Y} = \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_N}\}$ ($\boldsymbol{y_i}$ is a astronomical time series and N is the number of astronomical time series) for the given number of component frequencies m, nyquist frequency f_{nyq}, and domain feature $\mathcal{F}_d$. Astronomical time series $\boldsymbol{y_i}$ is the time sequence of magnitude data points for each celestial object. The component frequencies are $(f_{\mathrm{nyq}}/m, 2f_{\mathrm{nyq}}/m, \ldots, f_{\mathrm{nyq}})$. Periodic feature $\mathcal{F}_p$ is obtained by transforming each time series $\boldsymbol{y_i}$ into m components frequencies, which are calculated by nyquist frequency f_{nyq} (Section 3.2), and Domain feature $\mathcal{F}_d$ represents the specialized natures of each celestial object (Section 3.3). We describe the details of those features in the following sections.

3.2. Periodic feature

We employ sparse modeling to extract periodic feature $\mathcal{F}_p$ in order to take advantage of the inherent characteristic of variable celestial objects.

3.2.1. Variables design

In order to extract representative frequencies from sparse time series datasets, we first design explanatory variables $X \in R^{n\times m}$ in Eq. (2) by leveraging trigonometric interpolation (Makarchuk et al., 2022). Trigonometric interpolation is an interpolation method in mathematics using trigonometric polynomials, which is a finite linear combination of sine and cosine terms. Since those terms are periodic, this method is suited for the interpolation of time series. Also, note that sine and cosine terms with the same frequency in trigonometric polynomials are inseparably related.

In detail, by leveraging the trigonometric interpolation, we estimate the regression curve using the following polynomials from astronomical time series $\boldsymbol{y_i}$ with unevenly spaced astronomical time series due to missing values:

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}}(t) = \sum_j w_j \cos(2\pi t \nu_j) + w_{m+j} \sin(2\pi t \nu_j) \tag{4}$$

In this equation, $\nu_j = j f_{\mathrm{nyq}}/m$ is a component frequency where f_{nyq} nyquist frequency, which is the maximum component frequency, and w_j is their amplitudes. The amplitudes w_j and w_{m+j} are fit by a linear model, and they are composed into $\widetilde{w}_j$ as follows:

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}}(t) = \sum_j \widetilde{w}_j \sin(2\pi t \nu_j + \alpha_j) \tag{5}$$

Algorithm 1 Periodic Analysis

Input: i) explanatory variables: X
ii) target variables: $\boldsymbol{y}$
iii) group set: $\mathcal{G}$
Output: i) regression coefficient: $\tilde{\boldsymbol{w}}$
ii) coefficient of determination: reg_score

1: // Group Lasso
2: $\boldsymbol{w} = \boldsymbol{0}$ // initialization
3: **while** $\boldsymbol{w}$ is not converged **do**
4: $\hat{\boldsymbol{w}} \leftarrow \boldsymbol{w} - \eta \nabla f(\boldsymbol{w})$ // $\nabla f(\boldsymbol{w}) = X^\top (X\boldsymbol{w} - \boldsymbol{y})/n$
5: **for** $\boldsymbol{G_j}$ to $\mathcal{G}$ **do**
6: $\boldsymbol{w}_{\boldsymbol{G_j}} \leftarrow \max(0, 1 - \eta\lambda/||\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}||)\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}$
7: **end for**
8: **end while**
9: reg_score $= ||X\boldsymbol{w} - \overline{\boldsymbol{y}}||^2/||\boldsymbol{y} - \overline{\boldsymbol{y}}||^2$
10: $\tilde{\boldsymbol{w}} = [\sqrt{{w_1}^2 + {w_{m+1}}^2}, ..., \sqrt{{w_m}^2 + {w_{2m}}^2}]$
11: **return** $\tilde{\boldsymbol{w}}$, reg_score

where $\widetilde{w}_j = \sqrt{w_j^2 + w_{m+j}^2}$, and $\alpha_j = \arctan(w_j/w_{m+j})$.

In our approach, we set explanatory variables $X \in R^{n\times 2m}$ in Eq. (2) with m sine and m cosine terms (component frequencies) as follows:

$$x_{ij} = \begin{cases} \cos(2\pi t_i \nu_j) & (j \le m) \\ \sin(2\pi t_i \nu_j) & (j > m) \end{cases} \tag{6}$$

In order to ensure the inseparable relationship between sine and cosine terms with the same frequency in trigonometric polynomials, we employ Group Lasso as a sparse modeling technique to impose the inseparable relationship between explanatory variables and extract relevant explanatory variables from sparse time series. Specifically, we define group set $\mathcal{G}$[1] $= \{\{1, m+1\}, \{2, m+2\}, \ldots, \{m, 2m\}\}$ in Eq. (2), which groups sine and cosine terms with the same frequency.

3.2.2. Periodic feature

In this section, we describe the periodic feature $\mathcal{F}_p$ using the results of Group Lasso. $\mathcal{F}_p$ is composed of regression coefficients $\tilde{\boldsymbol{w}}$ and the coefficient of determination of regression *reg_score*, and it is used as input features of variable celestial object detection. The nature and rationale behind adopting these values are described below:

$\tilde{\boldsymbol{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_m)$

This vector is the coefficients of component frequencies. Each coefficient represents the importance of the corresponding frequency.

reg_score

This value is the coefficient of determination (R^2), a statistical measure of how accurately the input series is modeled by $\tilde{\boldsymbol{w}}$. We use this value as an input feature since if the accuracy of $\tilde{\boldsymbol{w}}$ is high, we can effectively detect variable celestial objects.

3.2.3. Algorithm

The pseudo-code of the periodic analysis is shown in Algorithm 1. In line 2, we initialize $\boldsymbol{w}$ with a zero vector. In lines 3–8, we update $\boldsymbol{w}$ according to the update formula of Group Lasso. In detail, $\nabla f(\boldsymbol{w})$ in line 4 indicates the direction of decreasing loss. In line 6, if $\|\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}\| > \eta\lambda$ holds, $\boldsymbol{w}_{\boldsymbol{G_j}}$ is a significant feature group for modeling. We repeat the above update until convergence and calculate the coefficient of determination reg_score of $\boldsymbol{w}$. In line 10, in accordance with the inseparable relationship between the sine and cosine terms with the same frequencies, $\boldsymbol{w}$ is transformed into $\tilde{\boldsymbol{w}}$ using Eq. (5). Finally, $\mathcal{F}_p$ is returned as a pair of $\tilde{\boldsymbol{w}}$ and *reg_score*.

[1] $\{j, m+j\} \in \mathcal{G}$ corresponds to the group of $\{\cos(2\pi t_i \nu_j), \sin(2\pi t_i \nu_j)\}$.

3.3. Domain feature

In addition to utilizing the periodic feature $\mathcal{F}_p$, we carefully design an additional periodic feature $\mathcal{F}_d$ for improving the accuracy of variable celestial object detection. Specifically, $\mathcal{F}_d$ is composed of the *average m_ap40*, *z_apertureflux_40_mag*, *diff_ap40* and *data_size* (Aihara et al., 2019). The nature and rationale behind adopting these values are described below:

average m_ap40

This average value is the mean of *m_ap40*. Each data point of *m_ap40* is the brightness (in magnitude) obtained within the circle of a 4-arcsecond diameter centered on the target celestial object in each exposure. The *m_ap40* is the target variable in the periodic analysis of the proposed model (the red data points in Fig. 1). Note that *m_ap40* may be affected by neighbor objects when the neighbor objects are bright and close to the target object as the measurements are performed on the image without deblending.

z_apertureflux_40_mag

This value is the sum of flux density obtained from the circle with a 4-arcsecond diameter centered on the target celestial object in a *Coadd image*[2]: *Coadd image* is generated by combining all CCD images of the celestial object. Since *Coadd image* has a higher signal-to-noise ratio and can avoid the effects caused by noises in each exposure images, *z_apertureflux_40_mag* is expected to be closer to the actual brightness of celestial objects than *average m_ap40*.

diff_ap40

This value is computed as the difference between *average m_ap40* and *z_apertureflux_40_mag*. Since *z_apertureflux_40_mag* is expected to have higher accuracy than *average m_ap40*, it represents the noise affecting rate of *average m_ap40*. Therefore, *diff_ap40* is useful in the detection of variable celestial objects.

data_size

This value is the number of observations. Although we attempt to mitigate overfitting by utilizing sparse modeling, overfitting is inevitable when *data_size* is small. So, this value can be utilized as the overfitting ratio.

Fig. 2 depicts the relationship between *diff_ap40* and *average m_ap40*. We can observe that celestial objects of low brightness have large *diff_ap40*. That means that the fainter celestial objects are more susceptible to noise by neighbor objects. Therefore, *diff_ap40* can be used as an effective domain feature expressing the accuracy degree of the periodic analysis result (both features computed from *m_ap40*). We achieve more accurate classification by utilizing not only the periodic pattern but also this insight.

3.4. Detection of variable celestial objects

We utilize two feature lists, $\mathcal{F}_p$ and $\mathcal{F}_d$ introduced in the previous sections, for the problem of detecting variable celestial objects (See Fig. 3 for details). We solve this problem by binary classification, either *objclass* is 0 (non-variable celestial objects) or 1 (variable celestial objects). We adopt LightGBM (Ke et al., 2017) to solve this task, which achieves state-of-the-art performances in many machine-learning tasks. However, the proposed method can adopt other binary classification approaches such as XGBoost (Chen and Guestrin, 2016).

[2] The resultant image made by statistically coadding each exposure images.

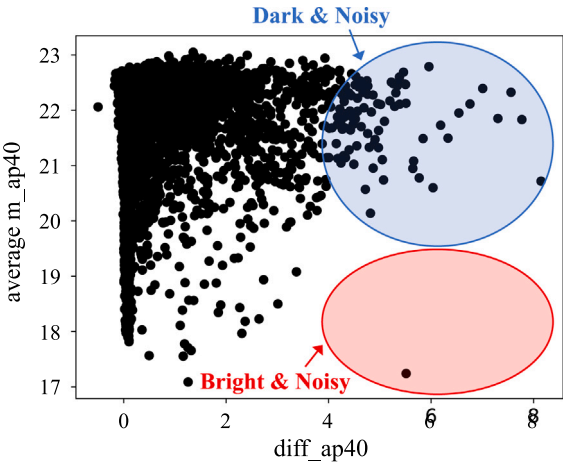


Fig. 2. The relationship between *diff_ap40* and *average m_ap40*: Note that *diff_ap40* corresponds to noise. This figure shows that almost all noisy celestial objects with large values of *diff_ap40* are relatively dark.

Algorithm 2 Detection of Variable Celestial Objects

Input: i) time series dataset of celestial objects: $\mathcal{Y}$
ii) number of component frequencies: m
iii) nyquist frequency: f_{nyq}
iv) domain feature: $\mathcal{F}_d$
Output: list of variable celestial object candidates: $\mathcal{L}$

1: // setting the variables
2: **for** $i = 1$ to n **do**
3: **for** $j = 1$ to m **do**
4: // $t_i = i, \quad \nu_j = j\frac{f_{\mathrm{nyq}}}{m}$
5: $X[i][j] \leftarrow \cos(2\pi t_i \nu_j)$
6: $X[i][j+m] \leftarrow \sin(2\pi t_i \nu_j)$
7: **end for**
8: **end for**
9: **for** $j = 1$ to m **do**
10: $\mathcal{G}[j] \leftarrow [j, j+m]$
11: **end for**
12: // Periodic analysis
13: **for** $i = 1$ to N **do**
14: $\mathcal{F}_p \leftarrow \mathrm{Algorithm1}(X, \mathcal{Y}[i], \mathcal{G})$
15: $\mathcal{F}[i] \leftarrow [\mathcal{F}_p, \mathcal{F}_d[i]]$
16: **end for**
17: // binary classification
18: $\mathcal{F} \rightarrow X_{train}, X_{test}, y_{train}, y_{test}$
19: lgb = LightGBM.train(X_{train}, y_{train})
20: $\mathcal{L}$ = lgb.predict(X_{test})
21: **return** $\mathcal{L}$

3.4.1. Algorithm

The pseudo-code of the overall proposed method is shown in Algorithm 2. In lines 2–11, we prepare the required variables. Specifically, we prepare explanatory variables X in lines 2–8, and the group set $\mathcal{G}$ in lines 9–11. In lines 13–16, we conduct periodic analysis(Algorithm 1) and prepare feature $\mathcal{F}_p$ using the results of the periodic analysis. Finally, in lines 18–20, we execute binary classification as described in this section.

3.5. Distributed processing

We leverage distributed servers for detecting variable celestial objects from large-scale datasets. We utilize distributed database Hive for storing the data points of celestial objects, and Apache Spark

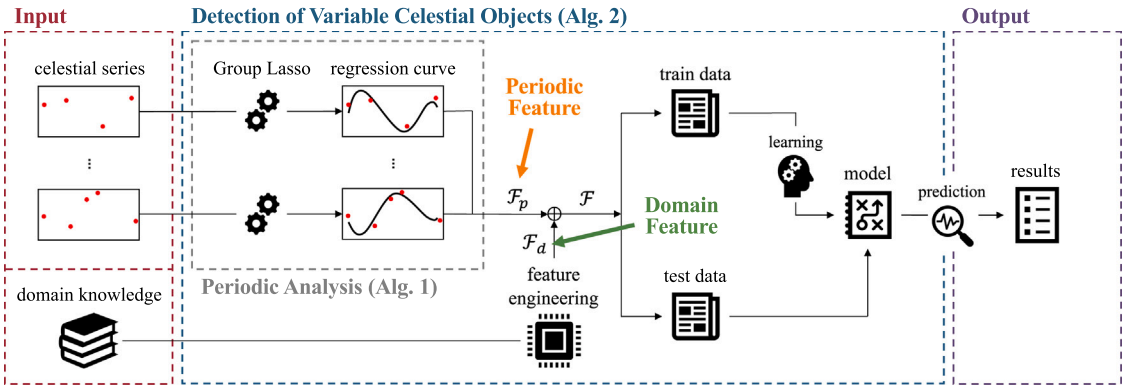


Fig. 3. Proposed method architecture: Periodic feature $\mathcal{F}_p$ is obtained by periodic analysis. Detection of variable celestial objects (blue block) detects variable celestial objects by utilizing both periodic feature $\mathcal{F}_p$ and additional domain feature $\mathcal{F}_d$.

distributed processing framework. Since Python is the most popular programming language for implementing machine learning techniques, we utilize SynapseML[3] that supports Python language for scalable machine learning pipelines on Apache Spark. In particular, we use the LightGBM implementation in SynapseML for detecting variable celestial objects and PySpark (Python API for Apache Spark)[4] for the periodic analysis.

4. Experiments

In this section, we evaluate the effectiveness of our proposed method. The questions we want to answer are as follows:

Q1 (effectiveness): How much more effective is the proposed method than a typical existing method, Sesar et al. (2007)? (Section 4.3)

Q2 (robustness): Is the periodic analysis in the proposed method robust against sparse astronomical time series? (Section 4.4)

Q3 (scalability): Does the proposed model scale well to dataset size? (Section 4.5)

Q4 (ablation study): Do both the sparse modeling and domain knowledge contribute to the model accuracy? (Section 4.6)

4.1. Experimental setup

We used the z band photometric catalog data for each exposure in the HSC PDR2[5] (Aihara et al., 2019), which are produced by forced photometry of each CCD images in the pipeline (Bosch et al., 2018), on the sky area of tract 9813. This dataset is composed of a celestial object identifier *objectid*, observation time *mjd*, magnitude *m_ap40* and *z_apertureflux_40_mag*, class label *objclass* (0 for non-variable and 1 for variable). We prepared the 5478 astronomical time series by sorting variable celestial objects based on the low-order statistics. The total data points are 393,534 (71.84 data points for each object on average). Furthermore, we manually assigned the class label *objectid* to the top 1557 astronomical time series in them sorted based on the statistics. As a result, 26.1% of 1557 astronomical time series are labeled as variable celestial objects, and all the rest (73.9%) are labeled as non-variable celestial objects. We refer to "26.1%" as a variable ratio. For our proposed model, we set the number of component frequencies $m = 500$, nyquist frequency $f_{\text{nyq}} = 1/30\,000$, penalty parameter $\lambda = 0.1$, and learning rate $\eta = 0.01$. We adopt this nyquist frequency for detecting variable celestial objects whose periodic length is between half a day and a few months.

[3] URL https://github.com/microsoft/SynapseML.
[4] URL https://spark.apache.org/docs/latest/api/python.
[5] https://hsc-release.mtk.nao.ac.jp/doc/.

Table 1
Evaluation metrics (AUC, accuracy, precision, recall, F1, MCC, Balanced Accuracy(BA)) scores of detection of variable celestial objects on PDR2.

Model	AUC	Acc.	Prec.	Rec.	F1	MCC	BA
Sesar et al.	0.750	0.743	0.507	0.507	0.507	0.334	0.667
Ours	**0.940**	**0.900**	**0.825**	**0.779**	**0.800**	**0.735**	**0.861**

4.2. Evaluation method

We used the method by Sesar et al. as the baseline, which identified variable objects using intrinsic variability σ, χ^2, etc. We choose this method because it is a popular method to detect variable celestial objects used by the National Astronomical Observatory of Japan. We utilized k-fold cross-validation to evaluate models. However, since this method is based on unsupervised learning, it cannot perform a label prediction in the same way as the proposed method, which is based on supervised learning. Therefore, in the baseline, we assumed that the ratio of variable celestial objects in the test dataset corresponds to the variable ratio in each fold. Specifically, in order to make a fair comparison, the baseline labeled the top 26.1% of the sorted variable by the statistics as variable celestial objects in each fold.

4.3. Q1. Effectiveness compared to the baseline

We evaluated the quality of the proposed method and the baseline, the method by Sesar et al. in terms of classification accuracy with respect to the seven evaluation metrics (AUC, accuracy, precision, recall, F1, MCC, Balanced Accuracy(BA)). Table 1 shows the proposed method outperforms the baseline in all metrics. This result indicates that we can extract the representative periodic patterns from extremely sparse astronomical time series by the sparse modeling technique and detect variable celestial objects by leveraging $\mathcal{F}_p$ and $\mathcal{F}_d$.

We also compare the proposed method and the baseline using confusion matrices in Fig. 4, in addition to the precision and recall we reported in Table 1. The false positive rate and false negative rate of the proposal method are very low, 0.003 and 0.005, respectively. In contrast, those of the baseline are relatively large, 0.297 and 0.059, respectively. This result also indicates that the proposal method outperforms the baseline method by Sesar et al.

In addition, we carried out another experiment using a balanced dataset to further validate the effectiveness of the proposed method. Specifically, we prepared the balanced dataset by randomly removing several astronomical time series that are labeled as non-variable celestial objects. The result shown in Table 2 verifies that the proposed method is capable of accurately classifying not only the imbalanced dataset but also the balanced one.

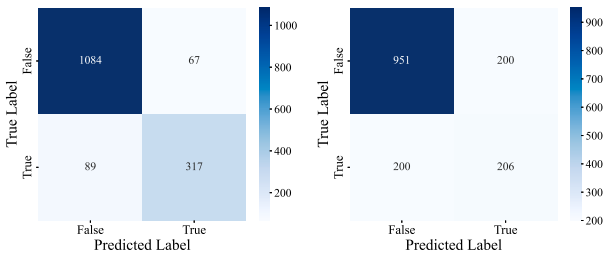


Fig. 4. Evaluation of detection of variable celestial objects on PDR2 using confusion matrices: The left and right matrices show the results of the proposed method and the method by Sesar et al. respectively.

Table 2
Evaluation metrics (AUC, accuracy, precision, recall, F1, MCC, Balanced Accuracy(BA)) scores of detection of variable celestial objects on the balanced dataset.

Model	AUC	Acc.	Prec.	Rec.	F1	MCC	BA
Sesar et al.	0.771	0.749	0.749	0.749	0.749	0.498	0.749
Ours	**0.931**	**0.888**	**0.926**	**0.842**	**0.882**	**0.780**	**0.888**

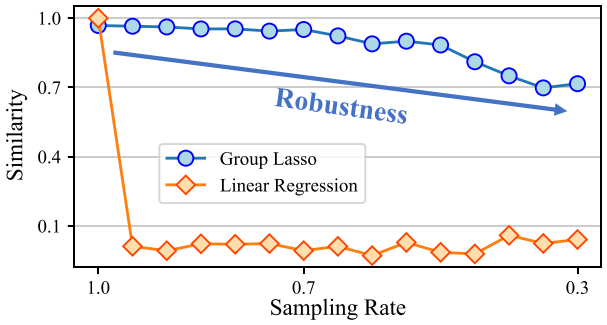


Fig. 5. Robustness of the proposed method: The x-axis denotes the sampling rate, namely the ratio of data points usage, while the y-axis denotes the similarity between the reconstructed series and the raw series. Group Lasso can reconstruct with fewer data points.

4.4. Q2. Robustness against sparsity

This experiment evaluated the robustness of our periodic analysis using the sparse modeling technique. If penalty parameter $\lambda = 0$, the Group Lasso would be equivalent to linear regression. Therefore, we compared the proposed approach to linear regression in the experiment. Note that linear regression does not perform the feature selection. We sampled a single astronomical time series in PDR2, *objectid* is "43158451320282758", which has a lot of data points (175). Fig. 5 shows the quality of the periodic analysis against various sampling rates. We used Cosine Similarity[6] Huang et al. (2008) for the quality evaluation metric to evaluate the error between the raw astronomical time series and the one obtained by each method because it is widely used for high-dimensional datasets. The x-axis in the figure, "Sampling Rate", represents the ratio of data points in the astronomical time series. The result shows that linear regression is susceptible to missing data; it cannot accurately analyze the astronomical time series even when the sampling rate is at 0.95. On the other hand, our periodic analysis is highly robust against low sampling rate, even when the sampling rate is at 0.3. Since the number of data points in the astronomical time series was 175, we expect that the proposed method can extract representative periodic patterns from the astronomical time series, whose number of data points is approximately 50.

[6] It measures the similarity in the direction of the two input vectors ignoring the scale by the cosine of the angle between them.

Table 3
Response time vs. input object size.

Input object size	900	1800	2700	3600	4500
Response time (s)	867	907	971	1014	1064

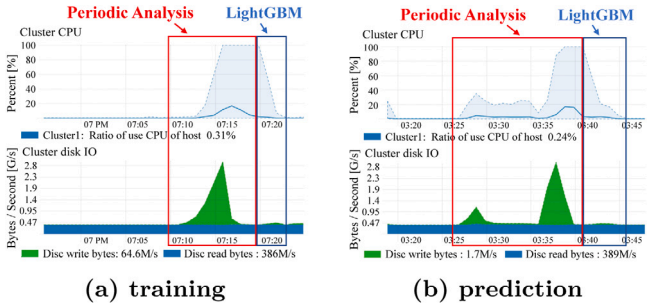


Fig. 6. CPU/disk IO performance profile for Spark cluster.

4.5. Q3. Scalability

We evaluated the scalability both on the periodic analysis step and the detection step of variable celestial objects using distributed processing framework (Apache Spark 3.2.0) and distributed database (Hive 1.1.0) using 66 computers (Ubuntu 16.04.5, Xeon processor Gold 6130 (2.10 GHz, 16core) × 2, memory 1.5 TB). We utilize PySpark 3.2.0 and Python 3.7.8 for the periodic analysis step and the LightGBM implementation in SynapseML 0.10.2 for detecting variable celestial objects. Table 3 shows how the response time changes when the number of input celestial objects increases. The response time is the summation of both steps of the periodic analysis and the detection of variable celestial objects. The results are approximated by $y = 0.0557x + 814$ using linear regression (correlation coefficient = 0.998), so the implementation is linearly scalable to the input size of celestial objects.

Fig. 6 depicts the Spark cluster performance of CPU utilization (%) and disk IO (GB/s) in the Y-axis during the training phase using labeled objects (left-hand side) and the prediction phase using non-labeled objects (right-hand side). X-axis indicates the elapsed time. In the CPU utilization figures (the upper part in Fig. 6), the solid lines show the average CPU utilization of all computers in the cluster and the dashed lines show the maximum CPU utilization in the computers. In the disk IO figures (the lower part in Fig. 6), the blue part shows the disk write IO and the green part show the disk read IO. Overall, we observe that (1) the periodic analysis occupies roughly 75% of the whole performance, (2) the periodic analysis step occupies 100% of the maximum CPU utilization, so it is the performance bottleneck of the whole system, and (3) the periodic analysis step also consumes large disk IO (2.9 GB/s at the maximum)

4.6. Q4. Ablation study

Table 4 shows the impact of the main components used in the proposed method using several variants of our models: without grouping, sparse modeling, periodic features, or domain knowledge features. Details about each component are described below:

Grouping: We conducted an experiment on a specific variant, namely *w/o. grouping*, to validate the effectiveness of the predefined groups of a group set $\mathcal{G}$ in Group Lasso. The *w/o. grouping* was derived using simple Lasso for the sparse modeling technique. We can discern from Table 4 that we enhance prediction accuracy at various evaluation metrics, except for Recall, by leveraging Group Lasso.

Sparse Modeling: We carried out an experiment on a specific variant, *w/o. sparse modeling*, in order to evaluate the effectiveness of the sparse

Table 4
Ablation study results: Averages and errors of evaluation metrics (AUC, accuracy, precision, recall, F1, MCC, Balanced Accuracy(BA)) calculated by prediction of multiple models. The evaluation method is k-fold validation (k = 10), and this error is based on standard deviation. The blue diamond is the mean value. These graphs indicate that (1) the complete model is more accurate than the others, (2) sparse modeling techniques and domain knowledge contribute to the accuracy improvement of the proposed model.

Model	AUC	Accuracy	Precision	Recall	F1	MCC	BA
Ours	$\mathbf{0.940 \pm 0.015}$	$\mathbf{0.900 \pm 0.012}$	$\mathbf{0.825 \pm 0.036}$	0.779 ± 0.047	$\mathbf{0.800 \pm 0.030}$	$\mathbf{0.735 \pm 0.035}$	$\mathbf{0.861 \pm 0.021}$
w/o. grouping	0.938 ± 0.020	0.898 ± 0.021	0.815 ± 0.052	$\mathbf{0.780 \pm 0.078}$	0.796 ± 0.057	0.729 ± 0.067	0.859 ± 0.039
w/o. sparse modeling	0.847 ± 0.031	0.821 ± 0.022	0.702 ± 0.063	0.538 ± 0.085	0.606 ± 0.065	0.503 ± 0.069	0.730 ± 0.040
w/o. $\mathcal{F}_d$	0.843 ± 0.038	0.812 ± 0.028	0.716 ± 0.053	0.468 ± 0.075	0.563 ± 0.060	0.468 ± 0.062	0.701 ± 0.034
w/o. $\mathcal{F}_p$	0.838 ± 0.044	0.821 ± 0.024	0.709 ± 0.069	0.533 ± 0.064	0.606 ± 0.057	0.504 ± 0.066	0.728 ± 0.034

modeling technique. As for the *w/o. sparse modeling* variant, it was achieved through the application of linear regression for periodic analysis. We find out in Table 4 that taking advantage of the sparse modeling technique is essential, as it improves the prediction accuracies.

Periodic Feature List $\mathcal{F}_p$: We executed an experiment on a variant identified as *w/o.* $\mathcal{F}_p$ with regards to the periodic pattern. It detect variable celestial objects solely through the exploitation of domain knowledge. As can be observed in Table 4, removing the domain knowledge leads to a significant performance drop. The inherent nature of variable celestial objects is critical for accuracy.

Domain Feature List $\mathcal{F}_d$: An experiment was performed on a variant referred to as *w/o.* $\mathcal{F}_d$ to assess the application of domain knowledge. It detected variable celestial objects by leveraging only periodic patterns. Inspection of Table 4 reveals a considerable decrease in performance when the periodic patterns were omitted. The specialized natures of variable celestial objects are also critical for accuracy.

5. Related work

The method by Sesar et al. (2007) is typical for the detection of variable celestial objects. It is a statistics-based method and differs from the proposed model. Specifically, it evaluates each celestial object by using some low-order statistics such as intrinsic variability σ, χ^2, average, standard deviation, and skewness of the magnitude of celestial objects, etc. Among the statistics in this method, intrinsic variability σ is the most dominant for the detection of variable celestial objects. Besides this method, there are several statistics-based methods that calculate statistics to detect variable celestial objects (Eyer et al., 2017; Shin et al., 2018). However, since the statistics are aggregations of astronomical time series, these methods fail to take advantage of the periodic information, which is the inherent nature of variable celestial objects.

The periodic analysis refers to the mining of periodic patterns. Namely, it is used to search for recurring patterns in time series. The Lomb–Scargle periodogram is a well-known algorithm for detecting periodic patterns in unevenly sampled time series and has been widely used within the astronomical community (Lomb, 1976; Scargle, 1982; VanderPlas, 2018). However, one notable issue with the Lomb–Scargle periodogram is the occurrence of aliasing. Aliasing is a phenomenon where certain frequencies can be confused with others, which can potentially complicate the analysis of periodicity.

6. Conclusion

In this paper, we proposed a method that detects variable celestial objects. Unlike existing methods that identify variable celestial objects using statistics, we approach detecting variable celestial objects by leveraging periodic information, which is one of their inherent characteristics. And since missing values in astronomical time series are too much, we utilize the sparse modeling technique that can train robust models by extracting relevant variables from sparse time series. Furthermore, our proposed model can efficiently process large-scale astronomical time series thanks to distributed processing. Experimentally, our proposed method shows outperforming the baseline in all five metrics. In addition, ablation study provided to testify the effectiveness of each component in our proposed method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This paper is based on results obtained from a project, JPNP16007, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Aihara, H., AlSayyad, Y., Ando, M., Armstrong, R., Bosch, J., Egami, E., Furusawa, H., Furusawa, J., Goulding, A., Harikane, Y., et al., 2019. Second data release of the hyper suprime-cam subaru strategic program. Publ. Astron. Soc. Japan 71 (6), 114.

Becker, A., Bochanski, J., Hawley, S., Ivezić, Ž., Kowalski, A., Sesar, B., West, A., 2011. Periodic variability of low-mass stars in sloan digital sky survey stripe 82. Astron. J. 731 (1), 17.

Bhatti, W.A., Richmond, M.W., Ford, H.C., Petro, L.D., 2010. Variable point sources in sloan digital sky survey stripe 82. I. Project description and initial catalog (0 hr α 4 hr). Astrophys. J. Suppl. Ser. 186 (2), 233.

Bosch, J., Armstrong, R., Bickerton, S., Furusawa, H., Ikeda, H., Koike, M., Lupton, R., Mineo, S., Price, P., Takata, T., et al., 2018. The hyper suprime-cam software pipeline. Publ. Astron. Soc. Japan 70 (SP1), S5.

Braga, V., Crestani, J., Fabrizio, M., Bono, G., Sneden, C., Preston, G., Storm, J., Kamann, S., Latour, M., Lala, H., et al., 2021. On the use of field RR lyrae as galactic probes. V. Optical and radial velocity curve templates. Astrophys. J. 919 (2), 85.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on KDD. pp. 785–794.

Eyer, L., Blake, C., 2005. Automated classification of variable stars for all-sky automated survey 1–2 data. Mon. Not. R. Astron. Soc. 358 (1), 30–38.

Eyer, L., Mowlavi, N., Evans, D., Nienartowicz, K., Ordonez, D., Holl, B., Lecoeur-Taibi, I., Riello, M., Clementini, G., Cuypers, J., et al., 2017. Gaia data release 1: The variability processing & analysis and its application to the south ecliptic pole region. arXiv preprint arXiv:1702.03295.

Gosnell, N.M., Gully-Santiago, M.A., Leiner, E.M., Tofflemire, B.M., 2022. Observationally constraining the starspot properties of magnetically active M67 sub-subgiant S1063. Astrophys. J. 925 (1), 5.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Huang, A., et al., 2008. Similarity measures for text document clustering. In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, Vol. 4. pp. 9–56.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.

Li, A., Shan, S., Gao, W., 2011. Coupled bias–variance tradeoff for cross-pose face recognition. IEEE Trans. Image Process. 21 (1), 305–315.

Liu, G., Huang, Y., Bird, S.A., Zhang, H., Wang, F., Tian, H., 2022. Probing the galactic halo with RR lyrae stars- III. The chemical and kinematic properties of the stellar halo. Mon. Not. R. Astron. Soc. 517 (2), 2787–2800.

Lomb, N.R., 1976. Least-squares frequency analysis of unequally spaced data. Astrophys. Space Sci. 39 (2), 447–462.

Makarchuk, A., Kal'Chuk, I., Kharkevych, Y., Kharkevych, G., 2022. Application of trigonometric interpolation polynomials to signal processing. In: 2022 IEEE 4th International Conference on Advanced Trends in Information Theory (ATIT). IEEE, pp. 156–159.

Ofek, E.O., Soumagnac, M., Nir, G., Gal-Yam, A., Nugent, P., Masci, F., Kulkarni, S.R., 2020. A catalogue of over 10 million variable source candidates in ZTF data release 1. Mon. Not. R. Astron. Soc. 499 (4), 5782–5790.

Scargle, J.D., 1982. Studies in astronomical time series analysis. II-statistical aspects of spectral analysis of unevenly spaced data. Astron. J. 263, 835–853.

Sesar, B., Ivezić, Ž., Grammer, S.H., Morgan, D.P., Becker, A.C., Jurić, M., De Lee, N., Annis, J., Beers, T.C., Fan, X., et al., 2009. Light curve templates and galactic distribution of RR lyrae stars from sloan digital sky survey stripe 82. Astrophys. J. 708 (1), 717.

Sesar, B., Ivezić, Ž., Lupton, R.H., Jurić, M., Gunn, J.E., Knapp, G.R., De Lee, N., Smith, J.A., Miknaitis, G., Lin, H., et al., 2007. Exploring the variable sky with the sloan digital sky survey. Astron. J. 134 (6), 2236.

Shin, M.-S., Chang, S.-W., Yi, H., Kim, D.-W., Kim, M.-J., Byun, Y.-I., 2018. Detecting variability in massive astronomical time-series data. iii. variable candidates in the superwasp dr1 found by multiple clustering algorithms and a consensus clustering method. Astron. J. 156 (5), 201.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

VanderPlas, J.T., 2018. Understanding the lomb–scargle periodogram. Astrophys. J. Suppl. Ser. 236 (1), 16.

Vivas, A.K., Martínez-Vázquez, C.E., Walker, A.R., Belokurov, V., Li, T.S., Erkal, D., 2022. Variable stars in the giant satellite galaxy antlia 2. Astrophys. J. 926 (1), 78.

Yan-Ke, T., Ning, G., Zhi-Kai, L., Hai-Lian, Y., Wen-Hui, D., 2022. Research on periodicity of single sector variable star of TESS space satellite. Chin. Astron. Astrophys. 46 (1), 63–81.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (1), 49–67.