STU MTF
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA

Time series anomaly detection in a cataclysmic variable using Support Vector Machine

Denis Benka[1], Andrej Dobrotka[1], Maximilián Strémy[1]

[1]Advanced Technologies Research Institute, Faculty of Materials Science and Technology in Trnava, Slovak University of Technology in Bratislava, Bottova 25, 917 24 Trnava, Slovakia

STU MTF
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA

Abstract

Quasi-periodic oscillations (QPO's) in cataclysmic variables (CV) can be very subtle and their confidence could be of a less significance than the reality is. In our observed object, MV Lyrae, we focus on such QPO's. We simulated the data according to Timmer & Koenig (1995) and estimated sigma intervals. Some known (not obvious) QPO's fell under 1-σ and therefore are not significant regarding this method. We propose and evaluate Support Vector Machine (SVM) models trained to identify those QPO's.

Our main goal is to testify whether the accuracy of QPO detection using machine learning methods is of a higher percentage than the significance of confidence levels obtained by the use of Timmer & Koenig's simulations.

Introduction

We analysed the light curve of a nova-like CV, MV Lyr. These systems show high accretion rate meaning they are in a bright state and good candidates to study accretion processes. We will draw our attention to a phenomenon occurring during the accretion process called flickering which does not have a strict period of occurrence. We simulated this phenomena and used the data as a training dataset with two categories. As a machine learning algorithm we chose to use a Support Vector Machine (SVM) trained on simulated numerical data with/without a QPO. The QPO's do not produce a single peak but rather a hump i.e., Lorentz profile, meaning the frequencies are not constant and manifest in an interval.

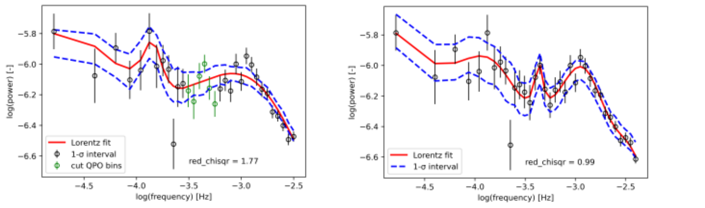

Fig. 1: Fitted simulated PDS and 1-σ intervals without the QPO of our interest (left), with the QPO (right). Source: own research.

Methodology

We analysed optical Kepler data of MV Lyrae, a nova-like system. We extracted a 272 days long light curve of the object. First, we divided the light curve. Using Lomb-Scargle (1976) (L-S) algorithm we subsequently calculated periodogram for each sample. The periodograms had a log-log scale. The main reason was the detection of QPO's. We constructed power density spectra (PDS) to see the variability more clearly. Bartlett's (1946) method was used for averaging the periodograms. Subsequently we created PDS with a binning of 0.05 bin length to average the 10-day periodograms. Obvious Lorentz profile is present at log(f) ≃ -3, while log(f) ≃ -3.4 shows only a subtle QPO (Fig. 1). As a next step we fitted the PSD with a Lorentz model consisting of four (or three) components. The fitting parameters were used as an input to simulations of the observed data. We used Timmer & Koenig (1995) method for simulations. We then created sigma intervals based on the simulations. The simulated PDS were used as a training dataset for our SVM in numerical form. One dataset consists of a 2D matrix (42x2) with frequency and the power of the signal in log scale. Several dataset sizes were used, i.e. 50, 100, 250, 500, 1000. Training with a validation set as well as k-fold cross validation were used. The kernel of the SVM was a radial basis function $\exp(-\gamma\|x-x'\|^2)$, where γ was optimised using a grid search as well as the margin parameter of the support vectors C of the SVM (see Tab. 1).

Tab. 1: Best-in-case training parameters with training accuracy. Source: own research

Validation		50	100	250	500	1000
Normal / k-fold	C	10/1	2/10	10/3	5	5
	γ	10/10	2/5	10/10	10	1
	acc	0.60/ 0.61	0.59/ 0.61	0.59/ 0.63	0.60/ 0.63	0.59/ 0.63

Results

For testing purposes, 1000 simulations from both categories as well as 26 real observational data with the QPO were used (12 with obvious QPO in PDS). Training accuracy was approx. 60% in all cases. Testing accuracy was approx. 45.5%. Increasing the size of the training dataset brought no significant changes to the testing and training accuracy. Using 10-fold cross validation increased the accuracy only by approx. 1%. Accuracy of testing with simulated data is 40%, accuracy of testing with real data is 51%. All results are summed in Tab. 2.

Tab. 2: Testing results of our SVM. Source: own research

Dataset size				50	100	250	500	1000
Testing accuracy	Normal cross validation	Simulated data	Without QPO	1	0	1	1	0
			With QPO	0	1	0	0	1
		Real data	Mixed	0.54	0.45	0.54	0.54	0.46
	K-fold cross validation	Simulated data	Without QPO	0	0	0	0	1
			With QPO	1	1	1	1	0
		Real data	Mixed	0.46	0.46	0.54	0.54	0.54

Discussion

Increasing the size of our dataset did not help the SVM to be more accurate. We tried a 5000-size dataset. Not only it trained for two weeks straight, but the results were even worse than the former (0.42 testing accuracy of real data). Even when we used grid search to find optimal hyperparameters C and γ, the accuracy remained unchanged. We are positive, that if we use our own kernel we could get more accurate results. On the other hand, we aim to use a Bayesian-based artificial neural network in our future research. Shapelet transform is a novel approach to anomaly detection (see e.g., Reich et al. 2019, Zhang et al. 2021, Wang et al. 2020) and we plan to combine it with the Bayes network. We first wanted to work with numerical data but we may transform it to image data in the future if we find it non-feasible because of the nature of the data.

Conclusions

We analysed the 272-day long Kepler light curve of MV Lyr. We performed time series analysis in order to confirm the existence of a QPO. We found a prominent one at log(f) ≃ -3. We were interested in one at log(f) ≃ -3.4 because of its subtle nature. Using sigma intervals from simulations, the latter fell into the 1-σ interval. We fitted the constructed PDS with a Lorentz profile. We simulated the PDS using the Timmer & Koenig (1995) method. We then used these numerical data as a training dataset. We divided it into two categories (with/without QPO). Out machine learning algorithm used was a SVM with a Radial basis function (Gaussian kernel) We used cross validation as well as k-fold cross validation. We tested with 1000 simulated data from each category as well as with 26 real data. Our testing accuracy was not plausible. In our future research we aim to use different methods for classification, maybe a different training data type.

Contact

Denis Benka
Advanced Technologies Research Institute
Bottova 25, 917 24 Trnava, Slovakia
@ denis.benka@stuba.sk
☎ +421 905 980 853

References

Bartlett, M. S., 1946. On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. Supplement to the Journal of the Royal Statistical Society, 8(1), pp. 27-41
Dobrotka, A., Negoro, H. & Konopka, P., 2020. Alternation of the flickering morphology between the high and low state in MV Lyr. Astronomy & Astrophysics, September, 641
Lomb, N., 1976. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science, Zväzok 39, pp. 447-462.
Reich, C., Nicolaou, C., Mansour, A., Van Laerhoven, 2019. Detection of Machine Tool Anomalies from Bayesian Changepoint Recurrence Estimation. 2019 IEEE 17th International Conference on Industrial Informatics, pp. 1297-1302
Timmer, J. & Koenig, M., 1995. On generating power law noise. Astronomy & Astrophysics, 300, pp. 707-710.
Wang, S., Zhang, S., Wu, T., Duan, Y., Zhou, L., Lei, H, 2020. FMDBN: A first-order Markov dynamic Bayesian network classifier with continuous attributes. Knowledge-Based Systems, 195
Zhang, H., Wang, P., Fang, Z. et al., 2021. ELIS++: a shapelet learning approach for accurate and efficient time series classification. World Wide Web 24, 511–539