Draft version May 1, 2020
Typeset using LaTeX **RNAAS** style in AASTeX62

Predicting Solar Flares Using Time Series Analysis

Lucas A. Pauker,[1] Monica G. Bobra,[2] and Eric Jonas[3]

[1] *Department of Physics, Stanford University, Stanford, CA 94305, USA*
[2] *W.W. Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305, USA*
[3] *Department of Computer Science, University of Chicago, Chicago, IL 60637, USA*

Keywords: Solar flares — Time series analysis — Machine learning

Rapidly emerging magnetic flux on the solar surface often indicates a greater likelihood of a solar flare (e.g. Leka & Barnes 2003). In this study, we attempt to answer the following question: Is there a characteristic pattern in the time evolution of magnetic flux or other physical parameters that distinguish flaring active regions from quiet ones?

In particular, we focus on deriving features that capture the shape, or trending behavior, of time series data. Others (e.g. Lee et al. 2018; Liu et al. 2019) successfully predicted flares by studying moments of time series data (such as the mean) or the last few points (by using a Long Short-Term Memory, or LSTM, network).

First, we assemble time series data for 17 different physical variables, such as the total magnetic flux, for every flaring active region observed between May 2010 and August 2019 (4500 in total). For a complete list of physical variables, see Table 3 of Bobra et al. (2014). We then segment these time series data into T-hour chunks (where T ranges from 2 to 24 hours) and ascribe each chunk into one of two classes: positive or negative. We define the positive class as T hours before a C, M, or X-class flare, and the negative class as T hours during a flare-quiet period.

Second, we extract features from these segmented time series data. In our model, we fit basis splines (also known as B-splines) to the data. These functions are defined below, where n is the order of the spline and t represents the number of knots or break points within the time series. We used $n = 3$ and $t = 11$, which produces 7 coefficients. We used these coefficients as features in our learning model.

$$\text{The base case } B_{i,0}(x) = \begin{cases} 1 & \text{if } t_i \leq x < t_{i+n} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Higher orders are defined recursively as } B_{i,k}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x)$$

Third, we use a machine learning algorithm for binary classification called a Support Vector Machine (SVM; Cortes & Vapnik 1995). An SVM works by first defining an N-dimensional feature space, where N is the number of features. In our case, each example (or flaring active region) includes 119 features (from time series data for 17 different physical variables, each characterized by 7 coefficients) and a label (positive or negative). The SVM then plots all the examples in the training data within this feature space. Finally, the SVM identifies a decision boundary within the feature space that best separates the positive examples from the negative ones. After training the SVM, we use it to predict the outcome of an unlabeled example. In other words, we use the SVM to predict whether a given T-hour chunk of time series data will culminate in a solar flare.

Our results are shown in Figure 1. We find that the B-spline fits of the active region area, total unsigned current helicity, total magnetic free energy, and the total unsigned flux time series best predict flaring activity. We also find that prediction accuracy increases with lag time, which indicates that several hours of time series data is necessary for

Corresponding author: Monica G. Bobra
mbobra@stanford.edu

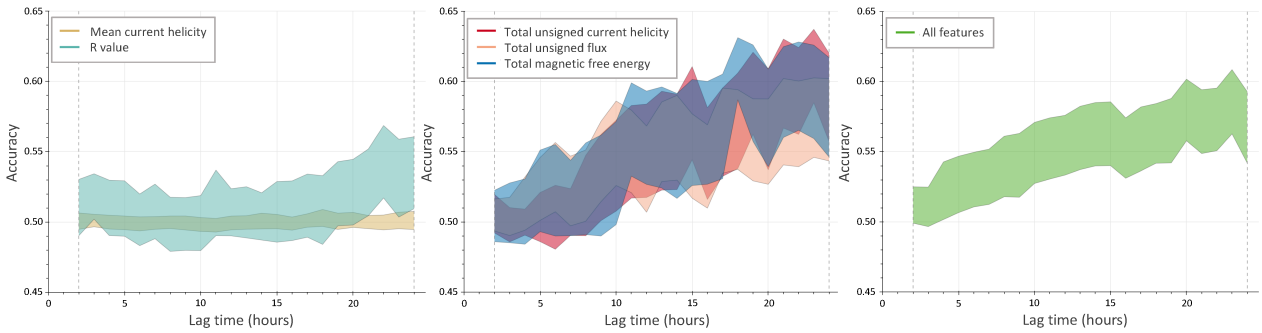

Figure 1. Plot of accuracy versus lag time for features that describe the B-spline fits of various time series. The width of the line represents the error in accuracy at each lag time. A lag time of zero represents the time of the flare.
Left: This plot shows no improvement in accuracy with increasing lag time using the mean current helicity and the R value (Schrijver 2007) time series.
Middle: This plot shows significant improvement accuracy with increasing lag time using the total unsigned current helicity, total unsigned flux, and magnetic free energy density time series.
Right: This plot shows some improvement in accuracy versus lag time using all 17 time series (with smaller errors).

maximizing predictive performance. We also find that combining features from multiple time series produces better predictive performance than using any individual time series.

This study represents a first step in time series analysis of flaring active regions. It is likely that physical variables on the Sun change states according to some unknown probabilities. Thus, using a multivariate state-space model may be a more robust approach. It is also likely that changing the definition of the negative class affects accuracy, and perhaps a 24-hour flare-quiet period is not sufficient to capture quiescent activity. We plan to address these issues in future work.

The data used here are courtesy of the GOES team and the *Helioseismic and Magnetic Imager* (HMI) science team of the NASA *Solar Dynamics Observatory.*

Software: A ll of the data and code used for this study are publicly available (Pauker et al. 2019). In addition, this study used the following open source software packages: AstroPy v3.1 (The Astropy Collaboration et al. 2018), Matplotlib v1.3 (Hunter 2007), NumPy v1.14.5 (van der Walt et al. 2011), Pandas v0.25.0 (McKinney 2010), scikit-learn v0.22.0 (Pedregosa et al. 2011), SciPy v1.2.2 (Jones et al. 2001), and SunPy v1.0.2 (SunPy Community et al. 2015) .

REFERENCES

Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, Solar Physics, 289, 3549, doi: 10.1007/s11207-014-0529-3

Cortes, C., & Vapnik, V. 1995, Machine Learning, 20, 273, doi: 10.1007/BF00994018

Hunter, J. D. 2007, Computing in Science and Engineering, 9, 90, doi: 10.1109/MCSE.2007.55

Jones, E., Oliphant, T., Peterson, P., et al. 2001. "http://www.scipy.org"

Lee, E.-J., Park, S.-H., & Moon, Y.-J. 2018, Solar Physics, 293, 159, doi: 10.1007/s11207-018-1381-7

Leka, K. D., & Barnes, G. 2003, The Astrophysical Journal, 595, 1296, doi: 10.1086/377512

Liu, H., Liu, C., Wang, J. T. L., & Wang, H. 2019, ApJ, 877, 121, doi: 10.3847/1538-4357/ab1b3c

McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, ed. S. van der Walt & J. Millman, 51 – 56

Pauker, L., Bobra, M., & Jonas, E. 2019, lucaspauker/hmi-time-series-analysis: Time Series Analysis with HMI SHARP data, 1.0, Zenodo, doi: 10.5281/zenodo.3384196

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Schrijver, C. J. 2007, ApJL, 655, L117, doi: 10.1086/511857

SunPy Community, Mumford, S. J., Christe, S., et al. 2015, Computational Science and Discovery, 8, 014009, doi: 10.1088/1749-4699/8/1/014009

The Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, The Astronomical Journal, 156, 123, doi: 10.3847/1538-3881/aabc4f

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science Engineering, 13, 22, doi: 10.1109/MCSE.2011.37