Astroinformatics
Proceedings IAU Symposium No. 325, 2017
M. Brescia, ed.

DOI: 00.0000/X000000000000000X

A Comparison of Classifiers for Solar Energetic Events

Graham Barnes[1], Nicole Schanche[2], K.D. Leka[1], Ashna Aggarwal[3], and Kathy Reeves[4]

[1]NorthWest Research Associates, 3380 Mitchell Ln, Boulder CO 80305, USA
email: graham@nwra.com, leka@nwra.com

[2]University of St Andrews, North Haught, St Andrews, Fife, KY16 9SS, UK
email: ns81@st-andrews.ac.uk

[3]Department of Earth, Planetary, and Space Science, University of California, Los Angeles CA 90095, USA
email: aaggarwal01@ucla.edu

[4]Harvard Smithsonian Center for Astrophysics, Cambridge MA 02138, USA
email: kreeves@cfa.harvard.edu

Abstract. We compare the results of using a Random Forest Classifier with the results of using Nonparametric Discriminant Analysis to classify whether a filament channel (in the case of a filament eruption) or an active region (in the case of a flare) is about to produce an event. A large number of descriptors are considered in each case, but it is found that only a small number are needed in order to get most of the improvement in performance over always predicting the majority class. There is little difference in performance between the two classifiers, and neither results in substantial improvements over simply predicting the majority class.

Keywords. methods: statistical, Sun: coronal mass ejections (CMEs), Sun: filaments, Sun: flares

1. Introduction

The Sun releases energy stored in its atmosphere by way of various pathways, including solar flares and filament eruptions. Predicting the occurrence of these events is important both for understanding the physical processes at work, and for mitigating the impacts at the Earth. Solar filaments are cool, dark channels of partially-ionized plasma that lie above the chromosphere. Their structure follows the neutral line between local regions of opposite magnetic polarity. The occurrence of filament eruptions is associated with coronal mass ejections (Schmieder et al. 2015; McCauley et al. 2015). A solar flare is a rapid, localized release of radiation, predominantly X-rays. They originate in the atmosphere above sunspot groups (active regions), where concentrations of strong magnetic field pass through the solar surface. Flares are often associated with CMEs and solar energetic particle events.

Previous investigations have shown at best modest improvements over simply always classifying as the majority class, despite a wide range of methods being considered, particularly for flares (e.g., Barnes et al. 2016). However, the different classifiers have typically been applied to different data sets, so it has been difficult to determine whether the ability to correctly classify these events is limited by the descriptors or the classifier. We present here a comparison of the performance of two types of classifier: Random Forest Classifier (RFC; Breiman 2001) and Nonparametric Discriminant Analysis (NPDA; Silverman 1986), applied to the data sets for two types of solar energetic events (filament eruptions and flares).

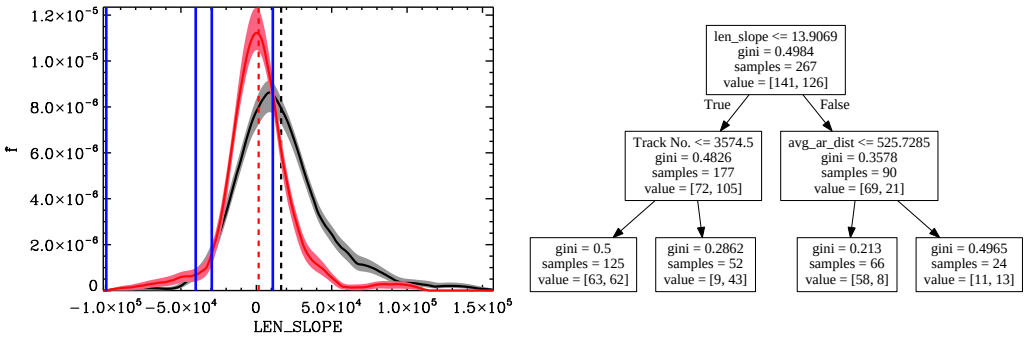

Figure 1. Left: Example nonparametric density estimates for eruptive (red) and quiescent (black) filaments for the rate of change of the filament length. Vertical blue lines show places where the density estimates are equal, while vertical dotted lines show the mean of each sample. Right: An example tree showing a split at the first level based on the value of the rate of change of the filament length.

2. Data and Classifiers

Filament Eruptions. For the filament eruptions, a total of thirty descriptors were used. The descriptors were computed for a sample of 126 filaments that erupted, and a sample of 141 that did not erupt from 2012-2013. The individual filament events were tracked and grouped together by an algorithm developed by Kempton & Angryk (2015). Most of these were filament features were taken from the Heliophysics Event Knowledgebase (HEK; Hurlburt et al. 2012) and characterize the morphology of the filament. This includes the length, area, tilt and number of barbs of the filament, and its chirality. For each feature, the minimum, maximum, mean, skew, and change in the value were computed. In addition, the decay index from 42-105 Mm above the filament channel was computed with the FORWARD code (Gibson et al. 2016).

Flaring Regions. For flaring active regions, a total of 412 descriptors were used. The descriptors were computed for a sample of 2623 HARPs (HMI Active Region Patches) that produced at least one C1.0 or larger flare within 24 hr and a sample of 25996 HARPs that did not produce any C1.0 or larger flare within 24 hr from 2010-2015. Most of the descriptors were characterizations of the photospheric magnetic field measured by the Helioseismic and Magnetic Imager (HMI; Hoeksema et al. 2014) on board NASA's Solar Dynamics Observatory (Pesnell et al. 2012). This included the first four moments of distributions, supplemented by totals where appropriate, of the different components of the field, the vertical current density, measures of the excess magnetic energy, properties of polarity inversion lines, the twist parameter, the shear and the current helicity (see Leka & Barnes 2007, for a list of descriptors). In addition, a simple coronal model based on magnetic charge topology (Barnes et al. 2005, and references therein) was computed. From this, additional descriptors characterizing the distribution of source properties and the connectivity matrix were computed (see Barnes & Leka 2006, for a list of descriptors). Finally, the past flaring history inferred from the peak GOES 1–8 Å flux was used to construct additional descriptors.

Classifiers. Two classifiers were applied to each data set: a Random Forest Classifier and Nonparametric Discriminant Analysis. Figure 1 shows an example of each for the filament eruption data. RFC is an ensemble method of machine learning that uses the aggregate of multiple decision trees to make classifications. Within a tree, each branch splits on a feature until the prediction is made. The tree shown has a depth of two, but for most of the results presented, the maximum depth was five. Ten-fold cross-validation

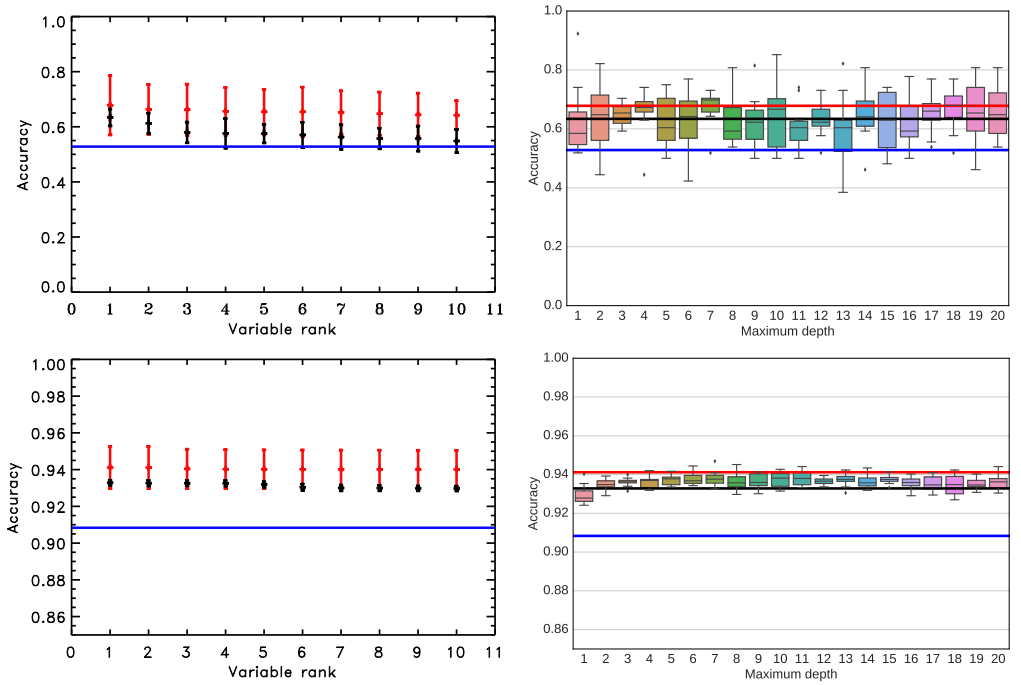

Figure 2. Accuracy of classifiers. The highest accuracy for one (black) and two (red) variable NPDA (left) and the accuracy as a function of the maximum depth of tree (with twenty trees in the forest) for the RFC (right) for filaments (top) and flares (bottom). The blue line in all panels is the accuracy from classifying as the majority class. Black and red lines in the right panels indicate the accuracy of the best performing descriptor(s) for one and two variable NPDA.

was used to remove bias and estimate uncertainties for the RFC. NPDA estimates the probability density of each population, in this case using the Epanechnikov kernel. A filament or active region is classified as belonging to the class with the higher density estimate at the value of its descriptor(s). For NPDA, cross-validation and a bootstrap were used to remove bias and estimate the uncertainties respectively.

3. Results

Figure 2 shows the accuracy (fraction of correct classifications) of the classifiers. Neither of the classifiers results in a substantial improvement in accuracy relative to always predicting the majority class, whether we consider filament eruptions or flares, and both produce classifications with similar accuracy. For NPDA, there is a small increase in accuracy, less than the estimated combined uncertainty, in using two descriptors instead of one. Similarly for the RFC, the accuracy does not significantly improve by increasing the depth or by adding more trees (not shown) beyond perhaps two or three. Thus only a small number of descriptors are needed to get the majority of the predictive power, which suggests that different descriptors have little independent information.

Figure 3 shows the descriptors that result in the highest accuracy classifications from NPDA and the most important descriptors in the RFC. There is considerable overlap, with NPDA and the RFC having approximately half of the top ten descriptors in common for both filament eruptions and flares. We therefore conclude that the limitation in predicting solar events is more likely to be the descriptors than the classifier.

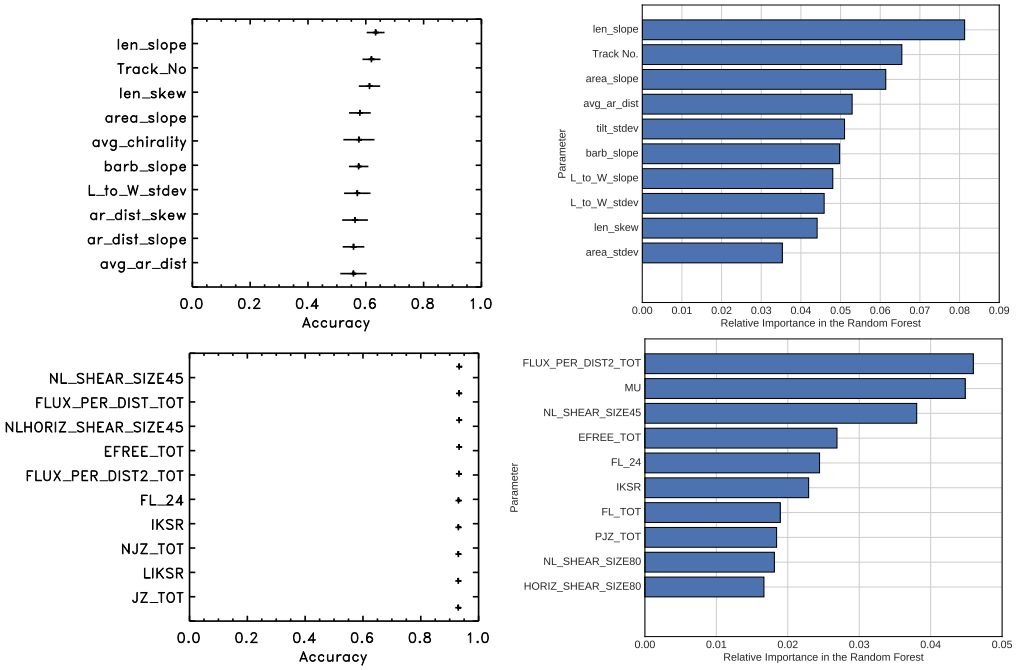

Figure 3. The most important descriptors. The best performing variables from one-variable NPDA based on the accuracy (left) and the relative importance of variables in the RFC (right) for filament eruptions (top) and flares (bottom).

Acknowledgements

This work began as a project during the 2015 Astro Hack Week, and has been partially supported by funding from the Division of Advanced Cyberinfrastructure within the Directorate for Computer and Information Science and Engineering, the Division of Astronomical Sciences within the Directorate for Mathematical and Physical Sciences, and the Division of Atmospheric and Geospace Sciences within the Directorate for Geosciences, under NSF awards #1443061 and #1630454.

References

Barnes, G. & Leka, K. D. 2006, *ApJ*, 646, 1303
Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, *ApJ*, 829, 89
Barnes, G., Longcope, D. W., & Leka, K. D. 2005, *ApJ*, 629, 561
Breiman, L. 2001, *Machine Learning*, 45, 5
Gibson, S., Kucera, T., White, S., et al. 2016, *Frontiers in Astronomy and Space Sciences*, 3, 8
Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, *Solar Phys.*, 289, 3483
Hurlburt, N., Cheung, M., Schrijver, C., et al. 2012, *Solar Phys.*, 275, 67
Kempton, D. J. & Angryk, R. A. 2015, *Astronomy and Computing*, 13, 124
Leka, K. D. & Barnes, G. 2007, *ApJ*, 656, 1173
McCauley, P. I., Su, Y. N., Schanche, N., et al. 2015, *Solar Phys.*, 290, 1703
Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, *Solar Phys.*, 275, 3
Schmieder, B., Aulanier, G., & Vršnak, B. 2015, *Solar Phys.*, 290, 3457
Silverman, B. W. 1986, Density Estimation for Statistics and Data Analysis (London: Chapman and Hall)