COLLOQUIA: SoHe3 2018

Flare forecasting and feature ranking using *SDO/HMI* data

M. PIANA([1])([3]), C. CAMPI([2]), F. BENVENUTO([1]), S. GUASTAVINO([1])
and A. M. MASSONE([1])([3])

([1]) *Dipartimento di Matematica, Università di Genova - Genova, Italy*
([2]) *Dipartimento di Medicina, Università di Padova - Padova, Italy*
([3]) *CNR SPIN Genova - Genova, Italy*

Summary. — We describe here the application of a machine learning method for flare forecasting using vectors of properties extracted from images provided by the *Helioseismic and Magnetic Imager* in the *Solar Dynamics Observatory* (*SDO/HMI*). We also discuss how the method can be used to quantitatively assess the impact of such properties on the prediction process.

1. – Introduction

Solar flares, the most explosive phenomena in the heliosphere, may extend to over $10,000$ km while releasing more than 10^{32} erg in less than 100 seconds, accelerating billions of tons of material to more than 10^6 km/h, emitting electromagnetic radiation at all wavelengths and, in this way, triggering the whole space weather connection. The full comprehension of solar (and stellar) flare physics is still an open issue, to such an extent that we can talk about a sort of flare paradox: simple computation based on their physical and geometrical properties and on magnetohydrodynamic (MHD) equations would lead to predict a light-up time for flares longer than 10^5 years, while the observed flash phase for these mysterious events is of the order of some minutes.

The numerical modelling of solar flare physics may rely on two different perspectives. On the one hand finite and boundary element methods applied against MHD partial differential equations allow the simulation of the electromagnetic fields and plasma properties in time and space; on the other hand, artificial intelligence allows pattern identification in the data mess and both source reconstruction with inverse methods and flare prediction with machine learning. In the present paper we will outline how a hybrid supervised/unsupervised machine learning method [1] is able to utilize features extracted from experimental observations of active regions.

2. – Setup of the problem

The data we consider in this paper are provided by the *Helioseismic and Magnetic Imager* in the payload of the *Solar Dynamics Observatory* (*SDO/HMI*). This telescope provides full disk vector magnetograms with a temporal cadence of 12 minutes, starting from february 2010. Relying on the *Solar Monitor Active Region Tracker (SMART)* tools and also using pattern recognition methods developed within the 'Horizon 2020' FLARECAST effort (http://flarecast.eu), the input data at disposal for the machine learning analysis become feature vectors of dimension up to 171 characterizing properties of the active regions (ARs) present in the *SDO/HMI* maps. Therefore the ingredients of our supervised approach are:

- A historical data set of feature vectors extracted from *SDO/HMI* data stored in the mission archive.

- A set of labels, each one associated to a feature set and encoding the outcome information, i.e. labels testifying the possible flare occurrence and intensity.

- A computational machine learning method that is trained by means of the historical (training) set and the corresponding set of labels.

When a new magnetogram arrives, the pattern recognition method extracts the features from it and the trained machine learning method both predicts the outcome corresponding to the new feature set and assesses the impact of each feature against the prediction effectiveness.

3. – The hybrid supervised/unsupervised machine learning method

The method we used for both prediction and feature ranking [1] is based on a hybrid supervised/unsupervised approach in which the supervised phase computes the weights with which each feature contributes to the prediction and the unsupervised phase provides an automatic, data-dependent threshold as a support for the binary prediction. More formally, we denote with X the set of N $F-$dimensional vectors representing the feature vectors at disposal in the training set. If y is the $N-$dimensional vector containing the observed binary labels, the supervised step is realized by means of the LASSO approach [2], which requires the solution of the minimum problem

$$(1) \qquad \hat{\beta} = \arg \min_{\beta}(\|y - X\beta\|_2^2 + \lambda\|\beta\|_1) \ .$$

The solution of this problem is the vector $\hat{\beta}$ of feature weights, with which we can forward-compute theoretical labels by means of

$$(2) \qquad \hat{y} = X\hat{\beta} \ .$$

Then, we applied a fuzzy clustering technique [3] on the computed labels in order to determine the centroids of the label-1 set (the flare occurs) and of the label-0 set (the flare does not occur). Therefore, when the new feature vector x_{new} arrives, $x_{new}^t\hat{\beta}$ is compared to the two centroids in order to decide the cluster to which it belongs. Further, the entries of $\hat{\beta}$ quantify the impact of each feature on the prediction task; therefore ranking them corresponds to rank the prediction impacts.

TABLE I. – *Skill scores obtained by the hybrid LASSO method described in this paper (HLA), a hybrid version of logit (HLO), a support vector machine for classification (SVC) and random forest (RF). TSS stands for true skill statistics, HSS for Heidke skill score, ACC for accuracy, FAR for false alarm ratio, POD for probability of detection. The numbers correspond to the average values over 100 realizations of the training set with the corresponding standard deviations.*

	TSS	HSS	ACC	FAR	POD
HLA	0.51 ± 0.04	0.54 ± 0.03	0.83 ± 0.01	0.29 ± 0.03	0.59 ± 0.04
HLO	0.50 ± 0.03	0.53 ± 0.03	0.83 ± 0.01	0.30 ± 0.03	0.59 ± 0.04
SVC	0.45 ± 0.03	0.50 ± 0.03	0.83 ± 0.01	0.25 ± 0.03	0.51 ± 0.04
RF	0.49 ± 0.03	0.52 ± 0.03	0.83 ± 0.01	0.28 ± 0.03	0.56 ± 0.03

4. – Results

We assessed the effectiveness of this approach for flare prediction and feature ranking by considering point-in-time *SDO/HMI* images in the time range between 09/14/2012 and 04/30/2016, with a time cadence of 24 hours. After the application of the pattern recognition step we had at disposal 4442 sets of $171-$dimension feature vectors. Then the training set was realized by randomly extracting 66% of the overall set of feature vectors and labeling each vector by annotating whether a flare with class at least $C1$ occurred in the next 24 hours (the same can be done for flares with class at least $M1$). The set of the remaining feature vectors were used as test set for the experiment. Table I explains the reliability of the prediction by comparing five standard skill scores [4] computed by means of the hybrid method and of other three machine learning methods [5], [6], [7]. Then, Table II shows how these same methods are able to rank the features according to their impact on the prediction.

5. – Conclusions

We showed how machine learning can be used for both binary flare forecasting and property ranking in the case of vector magnetograms provided by *SDO/HMI*. In particular, the ability to quantitatively assess the impact with which each property contributes to the prediction procedure has both physical and instrumental applications: high rank features are most likely associated to crucial physical processes and instruments that

TABLE II. – *Ranking of the properties extracted from ARs recognized in* SDO/HMI *magnetograms as provided by the same machine learning methods considered in Table I. The numbers in the first four columns correspond to rank values computed from the weight vectors (1) and averaged over 100 random realizations of the test set. The last two columns contain the average of the previous numbers over the four algorithm outcomes and the corresponding standard deviation, respectively.*

	HLA	HLO	SVC	RF	average	std
flare_index_past	13.98	28.84	19.91	3.51	16.56	10.63
sharp_kw/hgradbh/total	3.47	37	18.59	16.57	18.95	13.87
wlsg_br/value_int	3.74	14.43	22.86	43.14	21.04	16.68
sharp_kw/jz/max	26.05	28	16.94	18.58	22.27	5.29
sharp_kw/usiz/max	24.2	36.75	34.79	18.37	28.53	8.73

accurately observe properties with high rank are probably more worthwhile designing.

The hybrid method described in this paper is currently implemented in the platform developed within the 'Horizon 2020' FLARECAST effort, together with other 15 both supervised and unsupervised algorithms. This service is accessible for open download at http://flarecast.eu and is constantly updated also thanks to its highly modular design.

* * *

REFERENCES

[1] BENVENUTO F., PIANA M., CAMPI C. and MASSONE A. M., *Astrophys. J.*, **853** (2018) 90.
[2] TIBSHIRANI R., *J. R. Stat. Soc. B*, **58** (1996) 267.
[3] BEZDEK J. C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, (Kluwer, Norwell) 1981.
[4] BLOOMFIELD D. S., HIGGINS P. A., MCATEER R. T. and GALLAGHER P. T., *Astrophys. J.*, **747** (2012) L41.
[5] WU T. T., CHEN Y. F., HASTIE T., SOBEL E. and LANGE K., *Bioinformatics*, **25** (2009) 714.
[6] CORTES C. and VAPNIK V., *Mach. Learn.*, **20** (1995) 273.
[7] BREIMAN L., *Mach. Learn.*, **45** (2001) 5.