The Astrophysical Journal, 853:90 (9pp), 2018 January 20 https://doi.org/10.3847/1538-4357/aaa23c

A Hybrid Supervised/Unsupervised Machine Learning Approach to Solar Flare Prediction

Federico Benvenuto[1], Michele Piana[2], Cristina Campi[3], and Anna Maria Massone[3]
[1] Dipartimento di Matematica Università di Genova, via Dodecaneso 35, I-16146 Genova, Italy
[2] Dipartimento di Matematica Università di Genova and CNR—SPIN Genova, via Dodecaneso 35, I-16146 Genova, Italy; piana@dima.unige.it
[3] CNR—SPIN Genova, via Dodecaneso 33, I-16146 Genova, Italy
Received 2017 June 21; revised 2017 December 11; accepted 2017 December 13; published 2018 January 25

Abstract

This paper introduces a novel method for flare forecasting, combining prediction accuracy with the ability to identify the most relevant predictive variables. This result is obtained by means of a two-step approach: first, a supervised regularization method for regression, namely, LASSO is applied, where a sparsity-enhancing penalty term allows the identification of the significance with which each data feature contributes to the prediction; then, an unsupervised fuzzy clustering technique for classification, namely, Fuzzy C-Means, is applied, where the regression outcome is partitioned through the minimization of a cost function and without focusing on the optimization of a specific skill score. This approach is therefore hybrid, since it combines supervised and unsupervised learning; realizes classification in an automatic, skill-score-independent way; and provides effective prediction performances even in the case of imbalanced data sets. Its prediction power is verified against NOAA Space Weather Prediction Center data, using as a test set, data in the range between 1996 August and 2010 December and as training set, data in the range between 1988 December and 1996 June. To validate the method, we computed several skill scores typically utilized in flare prediction and compared the values provided by the hybrid approach with the ones provided by several standard (non-hybrid) machine learning methods. The results showed that the hybrid approach performs classification better than all other supervised methods and with an effectiveness comparable to the one of clustering methods; but, in addition, it provides a reliable ranking of the weights with which the data properties contribute to the forecast.

Key words: methods: data analysis – methods: statistical – Sun: flares – sunspots

1. Introduction

Solar flares are the most energetic events in the solar system. Over a typical duration of ∼(10–1000) s, they can release up to 10^{32} erg of energy—stored in stressed active region (AR) magnetic fields—into directed mass motions, heating, and acceleration of supra-thermal charged particles, including electrons, protons, and heavier ions (Kontar et al. 2011). Solar flares, together with coronal mass ejections, are the main drivers of space weather at Earth and can sometimes even significantly affect Earth- and space-based technology systems like power grids, flight navigation, and satellite communications (Balan et al. 2014; Hayes et al. 2016). Predicting solar flares requires, first of all, the determination of parameters such as properties of sunspot groups or of the coronal magnetic field configuration that are thought to be important for the understanding of fundamental processes in solar plasma physics. Second, at a more technological level, these parameters are used as input values for algorithms that realize predictions providing, for example (but not exclusively), a binary flare/no-flare outcome (Gallagher et al. 2002; Wheatland 2004; Bloomfield et al. 2012).

Most recent flare prediction algorithms belong to the machine learning framework (Li et al. 2007; Colak & Qahwaji 2009; Yu et al. 2009; Yuan et al. 2010; Bobra & Couvidat 2015). In this setting, data properties utilized for prediction are named *features*. In the case of *supervised* learning, a set of historical data is at out disposal where features are tagged by means of *labels* representing the observation outcome, and the prediction task consists of determining the label associated with the incoming features' set. On the other hand, *unsupervised* methods do not use any training set and data are clustered in different groups according to similarity criteria involving data features.

A crucial aspect of flare prediction, characterized by notable physical implications, is to provide hints on which data features mostly correlate with the labels. This information can be obtained by computing the feature weights and by ranking them according to their values. Methods that provide this kind of information are specific implementations of standard neural network approaches (Garson 1991; Olden et al. 2004) or by means of specific machine learning methods for a regression like LASSO (Tibshirani 1996), $l1$-penalized logit ($l1$-logit in the following; Wu et al. 2009), and random forest (RF in the following; Breiman 2001). Flare prediction with regression algorithms is typically obtained by accounting for numerical skill scores for the assessment of flare prediction performances (Bloomfield et al. 2012), focusing on one of them, and thresholding the regression outcome in such a way to optimize the selected score. The main drawback of this approach is that the thresholding process obviously depends on the skill score chosen for maximization: optimizing a specific score may result, and often does result, in poor values for the other scores.

The present paper introduces a novel approach to flare prediction, whose aim is to provide classification and feature weights computation in a completely automatic and skill-score-independent way. The perspective of such an approach is hybrid and rather general. First, a regularization method for regression is applied to the training set. This approach aims to optimize a function made of two terms: the discrepancy term measures the distance between prediction and data in the training set, while the penalty term (typically an $l1$ penalty

term) imposes a constraint on the number of features that significantly contribute to the prediction itself. More specifically, this regularization step reconstructs the vector of weights with which each feature contributes to the prediction in the training set. Then, the set of real values obtained by multiplying the weights times the feature values in the training set is automatically clustered in two classes by means of a *clustering* technique. Clustering is an unsupervised learning approach that organizes a set of samples into meaningful clusters based on data similarity. Data partition is obtained through the minimization of a cost function involving distances between data and cluster prototypes. Optimal partitions are obtained through iterative optimization: starting from a random initial partition, samples are moved from one cluster to another until no further improvement in the cost function optimization is noticed. Therefore, in the second step of the hybrid approach, clustering performs an automatic thresholding of the regression outcomes, which depends on the historical set used for the training phase (being, therefore, intrinsically data-dependent) and which is not based on tuning the values of a specific skill score (being, therefore, intrinsically skill-score-independent). The resulting algorithm presents several advantages with respect to standard one-step approaches: it selects the most significant features, since, in the first step, it relies on a regularization technique that promotes sparsity; it is a classification method, since at the end it produces two clusters, each one corresponding to a specific outcome of the prediction; it performs classification in a flexible, data-adaptive way, which makes it significantly efficient in providing good performances with respect to all standard skill scores. The hybrid approach in this paper utilized LASSO in the regularization step and Fuzzy C-Means (FCM; Bezdek 1981) to cluster the LASSO outcome. However, it is important to note that the first step could be in principle performed by any other regularization method for regression and the second step could be in principle performed by any other unsupervised clustering algorithm.

In order to corroborate the effectiveness of this hybrid approach we utilized a set of data from the National Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC) and compared our results with the ones provided by some of the most used machine learning approaches in flare forecasting. We found that the automatic classification provided by the hybrid method produces, on average, competitive results for all skill scores utilized in the paper.

The plan of the paper is as follows. Section 2 illustrates the kind of data that prediction algorithms will deal with. Section 3 introduces our hybrid approach for flare prediction with feature weights computation. Section 4 applies the hybrid approach to the set of SWPC data described in Section 2 and compares its performances to the ones obtained by other machine learning methods. Our conclusions are offered in Section 5.

2. SWPC Data

Solar ARs are classified according to magnetic field complexity indicators. For example, ARs tracked by the NOAA SWPC are typically classified by using the following five indicators (features): the area, three classes of the McIntosh indices (the Zurich class, the penumbral class, the compactness class; McIntosh 1990), and the Mount Wilson index (Hale et al. 1919). The area index is computed in fractions (millionths) of a solar hemisphere. The McIntosh scheme uses white light emissions to represent sunspot structure and is composed of three independent variables: the *Zurich class Z* of leading/trailing spot size and separation, which may assume seven categorical values; the *penumbral class p* of primary spot regularity, which may assume six categorical values; the *compactness class c* of internal spot distribution, which may assume four categorical values. Finally, the Mount Wilson scheme groups sunspots into classes based on the complexity of magnetic flux distribution in associated ARs, according to rules set by the Mount Wilson Observatory in California; this feature may assume eight categorical data. Therefore, in summary, each sample in the SWPC database is made of five features, four of which are categorical.

In order to apply machine learning algorithms, either supervised or unsupervised, we need to transform the categorical information contained in the above sunspot classifications (specifically, the McIntosh and Mount Wilson indices) into numerical data. This can be done by either transforming the categorical variables into *dummy variables* (Hardy 1993) or by computing occurrence frequencies in a historical database, i.e., by associating with each categorical variable the frequency with which a flare occurred or not in correspondence with that variable. In this paper, we used this second approach, which preserves the dimension of the space where to perform the data analysis (indeed, in this application, the use of dummy variables would increase the dimension of the data space up to 26). Specifically, we have considered the SWPC database covering the 1988 December to 1996 June time range, and we have computed the frequency with which a sunspot classified by a specific value of a fixed indicator produces a flare greater than a given class. Anyhow, we have verified that the use of the dummy variables does not improve the effectiveness of the prediction for all methods considered in this paper.

More formally, and focusing on the specific case of the value *A* for the Zurich class in the McIntosh classification (this value denotes one or more tiny spots that do not demonstrate bipolarity or exhibit penumbra), we denoted by $N_{Z=A}^{(C1\to C9)}$, $N_{Z=A}^{(M1\to M9)}$, and $N_{Z=A}^{(\geqslant X1)}$ the occurrences of flaring events of class *C*, *M*, and *X*, respectively, and computed the frequencies associated with flaring events of class greater or equal to a specific class as

$$f_{Z=A}^{(\geqslant C1)} = \frac{N_{Z=A}^{(C1\to C9)} + N_{Z=A}^{(M1\to M9)} + N_{Z=A}^{(\geqslant X1)}}{\#\,A \text{ occurrences}} \quad (1)$$

(with the corresponding no-flare-event frequency defined as $f_{Z=A}^{(\text{noflare})} := 1 - f_{Z=A}^{(\geqslant C1)}$);

$$f_{Z=A}^{(\geqslant M1)} = \frac{N_{Z=A}^{(M1\to M9)} + N_{Z=A}^{(\geqslant X1)}}{\#\,A \text{ occurrences}} \quad (2)$$

(with the corresponding no-flare-event frequency defined as $f_{Z=A}^{(\text{noflare})} := 1 - f_{Z=A}^{(\geqslant M1)}$);

$$f_{Z=A}^{(\geqslant X1)} = \frac{N_{Z=A}^{(\geqslant X1)}}{\#\,A \text{ occurrences}} \quad (3)$$

(with the corresponding no-flare-event frequency defined as $f_{Z=A}^{(\text{noflare})} := 1 - f_{Z=A}^{(\geqslant X1)}$). We note that, in these equations, $\#\,A$ occurences indicates the number of times in which the Zurich class assumes value *A* regardless of flare occurrence.

Similar formulas can be written for each one of the other categorical predictors.

We finally note that the frequencies computed according to the previously described process, together with the sunspot area, become the numerical features characterizing the input samples for all the machine learning methods used in this paper.

3. The Hybrid Approach

We introduce an approach to flare prediction with feature weights computation, which is hybrid, as it combines a supervised and an unsupervised algorithm, and skill-score-independent, as it performs classification without focusing on the optimization of a specific skill score.

We denote with X the matrix with dimension N (number of samples) $\times$ F (number of features) whose columns contain the feature values for each sample in the training set; β is the $F \times 1$ vector containing the F model parameters to determine and y is the $N \times 1$ data vector used in the training set and made of 0 and 1 values (where 0 means no flaring event and 1 indicates the flare occurrence). The first step of our hybrid two-step approach utilizes LASSO to compute feature weights. Specifically, we look for the solution of the minimum problem

$$\hat{\beta} = \arg\min_{\beta}(\|y - X\beta\|_2^2 + \lambda\|\beta\|_1), \tag{4}$$

where the regularization parameter λ is optimized by means of a cross-validation procedure (Stone 1974), $\|\cdot\|_2$ denotes the Euclidean norm, and $\|\cdot\|_1$ denotes the $l1$ norm. Then, in the second step, we apply a clustering method for partitioning $\hat{y}$, where $\hat{y} = X\hat{\beta}$. In a classical clustering approach like K-Means (Jain et al. 1999), each sample may belong to a unique cluster, while in a fuzzy clustering formulation a different degree of membership is assigned to each sample with respect to each cluster, which implies a much higher flexibility in accounting for data characteristics. Therefore, in the second step of our hybrid approach, we used FCM, which is the fuzzy extension of K-Means. In this framework, the FCM functional is given by

$$J_m(\hat{y}, \hat{z}, U) = \sum_{k=1}^{N}\sum_{j=1}^{2}(u_{jk})^m d_{jk}^2, \tag{5}$$

where $\hat{z} = \{\hat{z}_j | \hat{z}_j \in \mathbb{R}, j = 1, 2\}$ is the set of the two centroids of the two clusters, the component $u_{jk} \in [0, 1]$ of the $2 \times N$ matrix U represents the membership of the kth sample to the jth cluster, d_{jk} is the distance between the jth centroid and the kth sample, and m is the so-called fuzzifier parameter. The FCM optimization problem is the one to (iteratively) determine the components of the matrix U and of the vector $\hat{z}$ given the components of the vector $\hat{y}$ (this optimization problem is solved by means of a standard Picard iteration scheme; Bezdek 1981). When applied to the LASSO outcomes, this clustering procedure splits the regression values into two disjointed sets $\mathcal{Y}^+$ and $\mathcal{Y}^-$. Sorting the LASSO outcomes in ascending order, we have that $\mathcal{Y}^- = \hat{y}_1, \ldots, \hat{y}_t$ contains the smallest outcome values and $\mathcal{Y}^+ = \hat{y}_{t+1}, \ldots, \hat{y}_n$ contains the largest ones. In this way, we can define a skill-score-independent threshold as $t = (\hat{y}_{t+1} - \hat{y}_t)/2$ and the prediction function for a new sample x^{new} as 0 if $x^{\text{new}}\beta \leqslant t$ and 1 if $x^{\text{new}}\beta > t$.

In the next section, the performances of this hybrid approach to flare prediction are compared with the ones of seven standard supervised and unsupervised machine learning algorithms (we point out that all these methods are not hybrid, i.e., they are one-step approaches that are not combined in any way). Together with a single-step LASSO and a single-step FCM method, we will use K-Means clustering (Jain et al. 1999), $l1$-logit, a standard multilayer perceptron (MLP; Rumelhart et al. 1986), a support vector machine (SVM; Cortes & Vapnik 1995), and an RF algorithm (Breiman 2001). Concluding this section, we provide a quick review of the main mathematical aspects of these methods.

K-Means clustering is a simplified version of fuzzy clustering, in which the memberships u_{jk} are binary values (while in FCM they are real numbers between 0 and 1) and the functional to minimize has the form

$$J(\hat{y}, \hat{z}, U) = \sum_{k=1}^{N}\sum_{j=1}^{2} u_{jk} d_{jk}^2. \tag{6}$$

Also, this functional is minimized iteratively, in this case by means of a standard maximum likelihood algorithm (Duda & Hart 1973).

$l1$-logit has been designed "ad hoc" to perform classification with feature selection (Wu et al. 2009). This is a regularization method in which the discrepancy term relies on the assumption that the data components follow the Bernoulli distribution. Therefore, in $l1$-logit the optimization problem is

$$\hat{\beta} = \arg\min_{\beta}(\sum_{i=1}^{N}\log(\exp(-y_i(X_i^T w + c)) + 1) + \lambda\|\beta\|_1), \tag{7}$$

where c is a constant term to optimize and, as in the case of LASSO, the $l1$-norm penalty term constrains the number of features that significantly contribute to the prediction to be small.

MLP is by far the most common neural network model used in machine learning. The usual training algorithm, which is adopted in this application, is the error-back-propagation (Rumelhart et al. 1986). This is a gradient descent algorithm and uses a forward and a backward pass through the feedforward neural network. Then, the weights update is performed using the derivatives of the error function of the network with respect to the neural weights.

SVMs for classification are examples of regularized kernel methods, requiring the solution of the minimum problem

$$\hat{\beta} = \text{argmin}_{\beta} D(\beta), \tag{8}$$

where

$$D(\beta) = \frac{1}{2}\sum_{i,j}^{N} y_i y_j \beta_i \beta_j K(x_i, x_j) - \sum_{i}^{N}\beta_i, \tag{9}$$

subject to

$$\sum_{i}^{N} y_i \beta_i = 0 \quad 0 \leqslant \beta_i \leqslant C \;\forall\; i = 1, \cdots, N. \tag{10}$$

C is an upper bound and can be seen as a regularization parameter; x_i is a $1 \times F$ vector and represents the ith row of matrix X; and $K(x_i, x_j)$ is the kernel function, which, in our application, is a reproducing kernel Hilbert space.

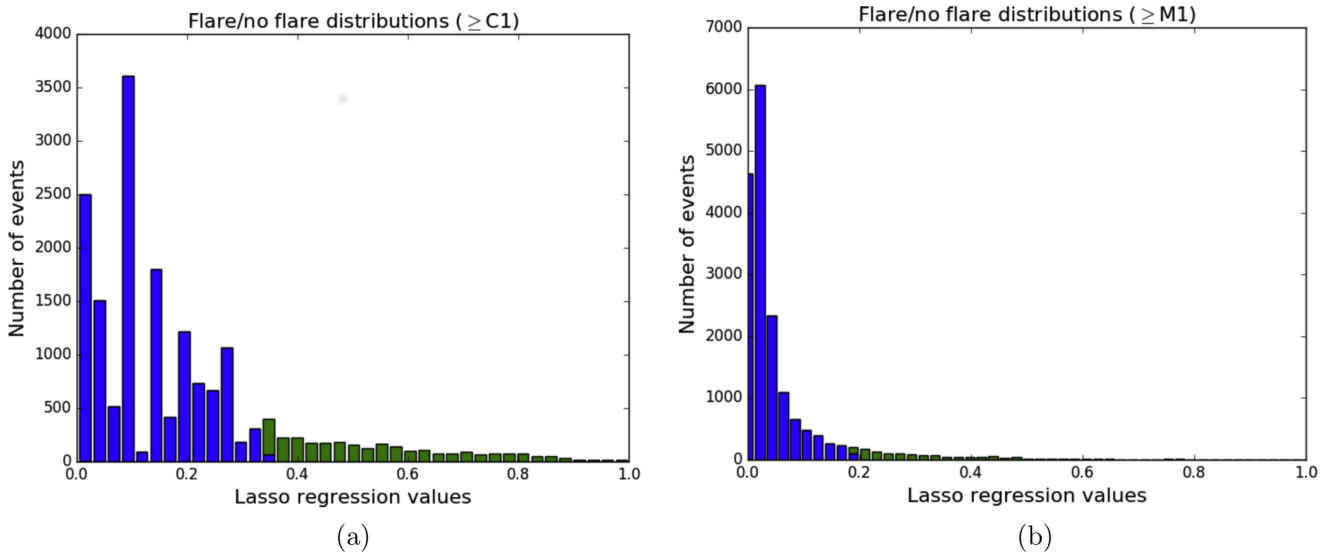

Figure 1. (a) $\geqslant C1$ class flare prediction. Split of the LASSO regression output by means of the Fuzzy C-Means algorithm. The x-axis shows the values of the regression outcomes provided by the cross validated LASSO algorithm. These values are binned with a bin value equal to 0.02. Blue and green colors represent the two clusters identified by the Fuzzy C-Means algorithm. Blue (green) cluster is the set of all the events for which the hybrid method returns a no-flare (flare) prediction. The only channel that is partly blue and partly green is the one where the threshold falls. This threshold value is 0.34. (b) The same as in (a), but for $\geqslant M1$ class flares. For this flare class the threshold value is 0.19.

RF is a rather novel machine learning technique, belonging to the family of the ensemble methods, i.e., methods that make use of a combination of different learning models to increase the classification accuracy. In particular, RF ensemble works as a set of decorrelated decision trees. Given a training set made of features/labels pairs, a decision tree classifier recursively splits training samples into subsets based on feature values. Each split is a node in the tree and the idea is to separate samples in the training set that have different characteristics by splitting the tree until every subset is made of samples belonging to the same class. Once the decision tree has been constructed, classifying a test feature is straightforward. Starting from the root node, one can apply the test condition to the record and follow the appropriate branch based on the outcome of the test.

4. Application to SWPC Data

In this section, we have validated the performances of our hybrid approach during the analysis of the SWPC test set covering the time range between 1996 August and 2010 December (the cardinality of such set is 22222; in this set, 17.6% of data corresponds to $\geqslant C1$ flares and 3.8% of data corresponds to $\geqslant M1$ flares); we used the data collected between 1988 December and 1996 June as training set (the cardinality of this second set is 17600; in this set, 19.6% of data corresponds to $\geqslant C1$ flares and 5.5% of data corresponds to $\geqslant M1$ flares). We have analyzed the same test set by means of seven other classical machine learning methods for classification: the (unsupervised) clustering K-Means and FCM algorithms, LASSO, $l1$-logit, a standard MLP (Rumelhart et al. 1986), an SVM (Cortes & Vapnik 1995), and an RF algorithm. For LASSO, $l1$-logit, SVM, MLP, and RF we used the same training set as in the case of the hybrid method. Further, both SVM and MLP have been used as classifiers as in Bobra & Couvidat (2015), i.e., events with a computed probability bigger than 0.5 are classified as flare, while events with a computed probability smaller than 0.5 are classified as no-flare. The same threshold has been used for $l1$-logit. On the other hand, RF is used in regression mode, and classification is obtained by averaging the RF probabilistic predictions (Pedregosa et al. 2011).

All of these prediction algorithms have been applied to predict flares with class above $C1$ and $M1$, respectively. From now on, for sake of brevity, we will indicate with $\geqslant C1$ and $\geqslant M1$ all flares with class above $C1$ and $M1$, respectively. We have not considered flares with class above $X1$ since they are very rare in this data set (less than 1% in the training set and around 0.5% in the test set).

Each sample is represented by a five-dimensional vector and, also thanks to the frequency calculation process described in Section 2, all components of such vectors are now real numbers. We have used for the analysis all five features introduced in Section 2, associated with NOAA/SWPC data. Note that the first four components range from 0 to 1, while the fifth one, i.e., the sunspot area, goes from 0 up to 10^2. Since the differences between component variances can affect the flare prediction performances, a standardization step preceded the application of the machine learning algorithms, i.e., each feature is transformed in such a way to obtain a variable with zero mean and unit variance. We also note that the frequency calculation must be performed for each case of interest, i.e., separately for the $\geqslant C1$- and $\geqslant M1$-flare predictions; therefore, for both the training set and the test set, we have constructed two subsets: the first subset, indicated with #1, is constructed using the frequencies of flares of class of at least C1 (i.e., by applying (1) and analogous); the second subset, indicated with #2, to the frequencies of flares of class of at least M1 (i.e., by applying (2) and analogous).

As explained in the previous section, the main advantage of the hybrid approach is in the fact that the way it partitions the set of LASSO outcomes is driven by the input data. This is clearly described in Figure 1, showing how FCM automatically identifies the probability threshold by automatically partitioning LASSO regression values in two classes.

Table 1
Confusion Matrices Corresponding to the Prediction of ⩾C1 Flares for All Methods Considered in the Paper

		Training		Testing	
Method		Actual True	Actual False	Actual True	Actual False
Hybrid	Predicted True	1739	1285	2087	1883
	Predicted False	1702	12874	1813	16435
Random Forest	Predicted True	1554	322	1442	1145
	Predicted False	1887	13837	2458	17173
Fuzzy C-Means	Predicted True	1650	1127	1972	1681
	Predicted False	1791	13032	1928	16637
K-Means	Predicted True	1626	1070	1946	1597
	Predicted False	1815	13089	1954	16721
LASSO	Predicted True	1063	450	1063	450
	Predicted False	2378	13709	2378	13709
*l*1-logit	Predicted True	1172	554	1488	887
	Predicted False	2269	13605	2412	17431
Multilayer Perceptron	Predicted True	1263	524	1584	1003
	Predicted False	2178	13635	2316	17315
Support Vector Machine	Predicted True	1157	489	1611	1028
	Predicted False	2284	13670	2289	17290

Table 2
Confusion Matrices Corresponding to the Prediction of ⩾M1 Flares for All Methods Considered in the Paper

		Training		Testing	
Method		Actual True	Actual False	Actual True	Actual False
Hybrid	Predicted True	478	953	456	1607
	Predicted False	494	15675	397	19758
Random Forest	Predicted True	420	47	150	429
	Predicted False	552	16581	703	20936
Fuzzy C-Means	Predicted True	589	1591	522	2284
	Predicted False	383	15037	331	19081
K-Means	Predicted True	527	1306	490	1954
	Predicted False	445	15322	363	19411
LASSO	Predicted True	122	59	122	59
	Predicted False	850	16569	850	16569
*l*1-logit	Predicted True	117	56	107	81
	Predicted False	855	16572	746	21284
Multilayer Perceptron	Predicted True	199	88	138	189
	Predicted False	773	16540	715	21176
Support Vector Machine	Predicted True	180	84	132	142
	Predicted False	792	16544	721	21223

The threshold value determines the prediction, whose performance can be measured by means of specific scores. Many skill scores can be found in the literature for the assessment of flare prediction performances (Bloomfield et al. 2012). All of these scores are linked to the forecast contingency tables made up of four elements:

1. The number of flares predicted and observed (true positives, TPs).
2. The number of flares not predicted but observed (false negatives, FNs).
3. The number of flares predicted but not observed (false positives, FPs).
4. The number of flares not predicted and not observed (true negatives, TNs).

These tables are known as *confusion matrices*, and we have computed them in Tables 1 and 2 in the case of the prediction of both ⩾C1 and ⩾M1 flares.

We have validated the eight flare prediction algorithms by means of the following skill scores defined in terms of the above elements. Specifically, the probability of detection (also

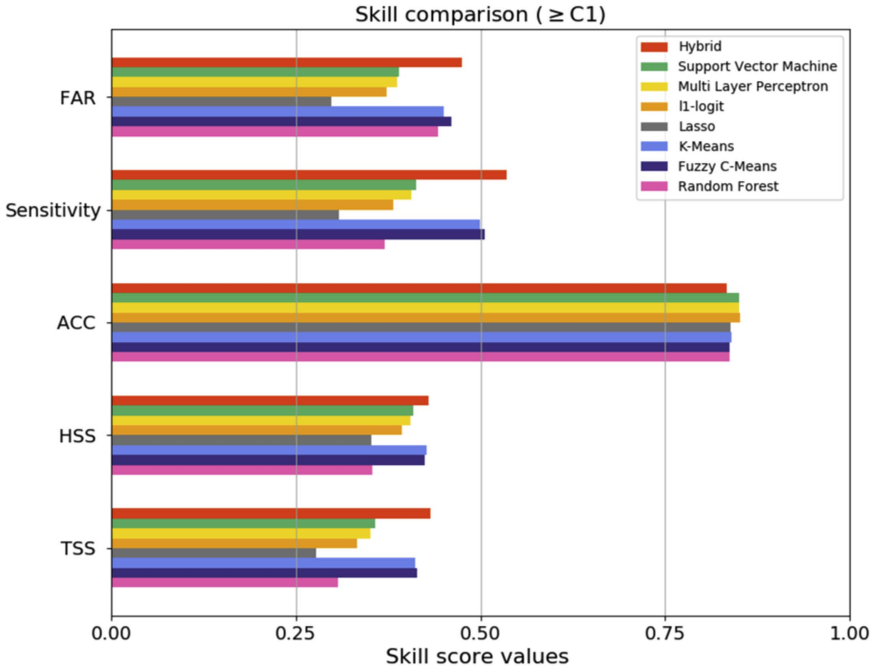

Figure 2. Comparison of performance between the eight flare prediction algorithms in terms of skill scores. The bar plots represent the skill score values obtained by applying each method to the test set for the prediction of $\geqslant C1$ flares.

called sensitivity)

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}\ ; \tag{11}$$

the accuracy

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}\ ; \tag{12}$$

and the false-alarm ratio

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}. \tag{13}$$

These scores range from 0 to 1 and the best predictions correspond to small FAR values and high values for the other scores. We also utilized two scores with values ranging from −1 to 1: the Heidke skill score

$$\mathrm{HSS} = \frac{2 \cdot (\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FN} \cdot \mathrm{FP})}{(\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{FN} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{FP} + \mathrm{TN})}\ , \tag{14}$$

and the true skill statistics

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{15}$$

Also, in this case, good prediction performances correspond to high values of the scores. Figures 2 and 3 present the values of all five skill scores for the $\geqslant C1$ flare prediction and the $\geqslant$M1-flare prediction, respectively. Moreover, Tables 3 and 4 provide the results of the processes feature weights computation performed by $l1$-logit, RF, and the hybrid technique. Specifically, the tables contain the weights β with which the sunspot area, the McIntosh indices, and the Mount Wilson index contribute to the flare prediction process for the three methods. Further, in order to evaluate the data selection dependence for training and testing on skill scores, we have applied a k-fold cross-validation (Burman 1989), where all available data are clustered into k subsets with the same cardinality; one of this subset is used as test set while the other ones are used as training sets, and the whole process is repeated k time with $k = 10$ (the balance between training and test sets is maintained during the k-fold cross-validation). The results of this procedure are represented in Figures 4 and 5.

We finally point out that all codes utilized in this analysis are available at the Github link https://github.com/midagroup/swpc, together with all optimized parameters.

5. Discussion and Conclusions

This paper introduces a novel approach to flare prediction, which utilizes indices associated with AR data and which is also able to automatically indicate the ones, among such features, that mostly contribute to the prediction. The approach is intrinsically hybrid, in the sense that it is based on the combination of the ability of regularization to compute feature weights with the ability of clustering to classify in a data-adaptive fashion. In the present implementation we have used LASSO in the step of feature weights computation and FCM in the clustering step. In fact, LASSO guarantees a notable degree of generality in regularization while FCM guarantees a notable degree of flexibility in data adaptation. It is interesting to note (see Figure 1) that FCM provides a threshold for the regression values different than the trivial 0.5 value.

Anyhow, in principle, any other kind of regression method could be used in the first step and any other kind of clustering method could be used in the second step. Actually, we have tested the hybrid approach using different combinations of the regression and clustering algorithms considered in this paper and found out that in all cases the use of clustering after

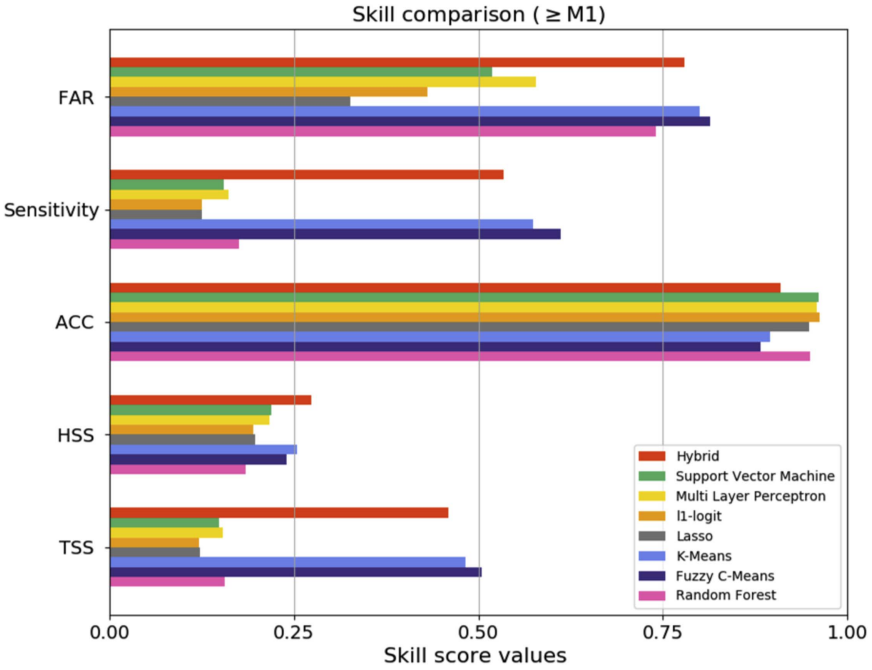

Figure 3. Same as in Figure 2, but for the prediction of $\geqslant M1$ flares.

Table 3
Feature Weights in $\geqslant C1$ Class Flare Prediction Computed from the Training Set

	MtWilson	McIntosh Z	McIntosh p	McIntosh c	Area
hybrid	0.198	0.268	0.222	0.164	0.147
$l1$-logit	0.189	0.332	0.219	0.154	0.104
RF	0.103	0.196	0.059	0.103	0.536

Note. For each method, the values correspond to the weights associated with the features divided by the sum of the weights, i.e., $\bar{\beta}_k := \hat{\beta}_k / \sum_{j=1}^{F} \hat{\beta}_j$.

Table 4
Feature Weights in $\geqslant M1$ Class Flare Prediction Computed from the Training Set

	MtWilson	McIntosh Z	McIntosh p	McIntosh c	Area
hybrid	0.281	0.117	0.047	0.146	0.407
$l1$-logit	0.181	0.264	0.217	0.142	0.194
RF	0.148	0.068	0.064	0.087	0.630

Note. See the caption of Table 1 for the meaning of the table entries.

regression improves the classification outcome on average over all skill scores.

We validated the approach against a NOAA SWPC data set and by comparing the results with the ones provided by standard machine learning flare prediction algorithms. Figure 2 shows that the hybrid approach does better than all other methods in predicting $\geqslant C1$ flares in the case of HSS and TSS that are often considered (Bloomfield et al. 2012) the most reliable skill scores in the game (for example, ACC tends to reach its maximum when the threshold is 0.5, which is not fully appropriate in the case of infrequent events such as M- and X-class flares). Anyhow, it also performs very well in the case of FAR and sensitivity, and its accuracy is comparable to the one of all other methods. In the case of the prediction of $\geqslant M1$ flares (see Figure 2), the hybrid method is still the most effective one if HSS is used and obtains a performance comparable to the ones of the unsupervised methods in the case of all other skill scores. More in general, the hybrid method predicts with a performance rate that is similar to the one of the two unsupervised clustering algorithms applied alone for all skill scores employed, and does almost systematically better than the supervised methods. This is reasonable, since the hybrid method utilizes FCM for clustering the outcomes of LASSO regression. However, while producing similarly competitive results in prediction, the hybrid method also provides feature weights computation, which is not the case for the unsupervised clustering methods. The k-fold cross-validation procedure in Figures 4 and 5 provides rather similar results: the performances of all methods are rather stable with respect to data randomization and, further, the hybrid approach is competitive with unsupervised clustering in flare prediction for almost all skill scores, while, at the same time, it provides feature weights ranking.

The higher forecasting effectiveness of the hybrid approach with respect to $l1$-logit, LASSO, SVM, RF, and MLP is due to the fact that it performs classification with a thresholding procedure that is data-adaptive, while here $l1$-logit, LASSO, SVM, and MLP utilize a fixed threshold. RF provides competitive results when the accuracy ACC score is considered. On the other hand, its performances with respect to HSS and TSS are rather similar to the ones provided by SVM and MLP, as already found in other applications (Wainberg et al. 2016). In general, we note that the threshold in $l1$-logit, SVM, and MLP could be tuned heuristically, searching for the values that provide the maximum for some specific skill scores. The main quality of the hybrid approach is just that its fuzzy clustering step realizes "a priori" and in an automatic way the classification of the regression outcomes; then, once the skill scores corresponding to this classification are computed "a

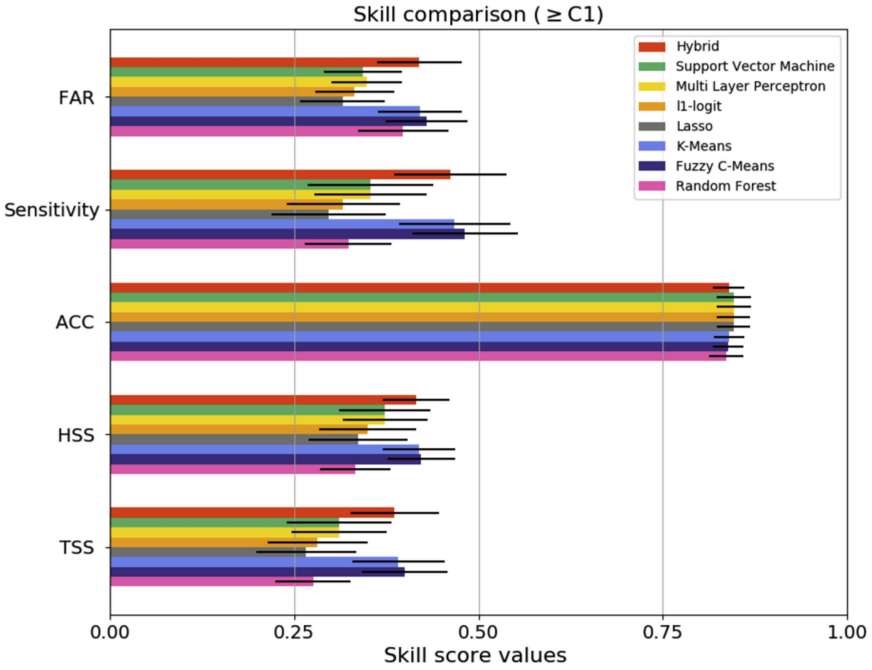

Figure 4. Averaged values and standard deviations of the same scores as in Figure 2 after a k-fold cross-validation procedure with $k = 10$ for the prediction of $\geqslant C1$ flares.

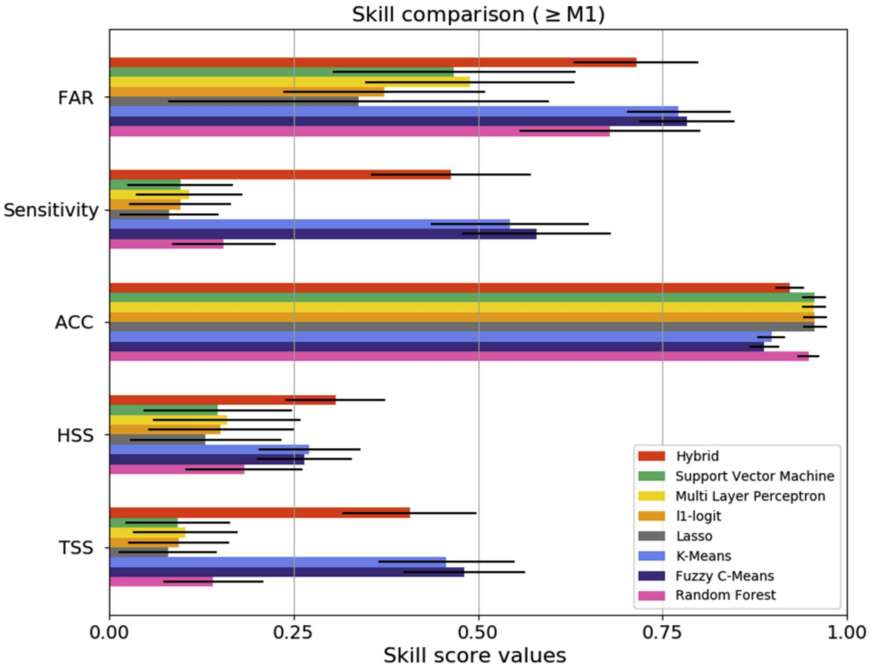

Figure 5. Same as in Figure 4, but for the prediction of $\geqslant M1$ flares.

posteriori," one realizes that this approach is competitive with all other methods with respect to all skill scores and, in particular, to TSS and HSS. For the sake of completeness, we compared TSS and HSS provided by the hybrid method with the ones of SVM, RF, and MLP when optimized with respect to such scores and found that:

1. If TSS is optimized, MLP provides a TSS value comparable with the one given by the hybrid method, while all other algorithms provide significantly worse performances. At the same time, HSS values either do not change or decrease with respect to the ones provided by using a fixed threshold.
2. If HSS is optimized, both HSS and TSS improve for all methods but not to the point to be competitive with the performances provided by the hybrid method.

We finally point out that the prediction power of all these machine learning methods significantly depend on the nature of the data utilized for training and testing. For example, SVM

provides TSS values greater than 0.7 when applied against *SDO*/HMI data, which are characterized by an amount of features and information significantly higher than the information content hidden in NOAA SWPC data (Bobra & Couvidat 2015; Muranushi et al. 2015; Nishizuka et al. 2017).

The hybrid approach, $l1$-logit, and RF also can be compared as far as their feature selection power is concerned. We note that, in this application, the number of features is very limited, and therefore these methods do not realize feature selection, but rather a ranking of the weights with which the different features impact the prediction process. If the input data set contains many features, like in the case of the features that can be extracted from *SDO*/HMI images, the feature selection procedure will set to zero all features useless for prediction. Table 3 clearly shows that, in forecasting $\geqslant$C1 flares, the hybrid method and $l1$-logit indicate the same features as the ones that mostly contribute to the prediction, while RF gives a lot of emphasis to the area. Results are different when predicting $\geqslant$M1 flares (see Table 4): both $l1$-logit and the hybrid method more strongly favor the AR area with respect to what happens in $\geqslant$C1 flares. On the other hand, the hybrid method gives greater emphasis to the Mount Wilson index. Interestingly, in the case of $\geqslant$M1 flares, RF provides the same feature ranking offered by the hybrid method. A clarification of this outcome shall be obtained by means of a systematic application of these two methods against either several SWPC data sets or features extracted from *SDO*/HMI images; this activity is part of the tasks currently addressed by the H2020 project FLARECAST, which will provide a technological platform for the testing of flare prediction algorithms and for the validation of the forecasting and feature weights computation results.

The authors have been supported by the H2020 grant Flare Likelihood And Region Eruption foreCASTing (FLARECAST), project number 640216. The authors kindly thank Prof. Shaun Bloomfield for providing the SWPC data and Dr. Annalisa Perasso for useful discussions.

ORCID iDs

Federico Benvenuto https://orcid.org/0000-0002-4776-0256
Michele Piana https://orcid.org/0000-0003-1700-991X
Cristina Campi https://orcid.org/0000-0003-2105-8554
Anna Maria Massone https://orcid.org/0000-0003-4966-8864

References

Balan, N., Skoug, R., Tulasi Ram, S., et al. 2014, JGRA, 119, 10041
Bezdek, J. C. 1981, Pattern Recognition with Fuzzy Objective Function Algorithms (Norwell, MA: Kluwer Academic)
Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJL, 747, L41
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Breiman, L. 2001, Mach. Learn., 45, 5
Burman, P. 1989, Biometrika, 76, 503
Colak, T., & Qahwaji, R. 2009, SpWea, 7, S06001
Cortes, C., & Vapnik, V. 1995, Mach. Learn., 20, 273
Duda, R. O., & Hart, P. E. 1973, Pattern Classification and Scene Analysis (New York: Wiley)
Gallagher, P. T., Moon, Y. J., & Wang, H. 2002, SoPh, 209, 171
Garson, G. D. 1991, Artif. Intell. Expert, 6, 47
Hale, G. E., Ellerman, F., Nicholson, S. B., & Joy, A. H. 1919, ApJ, 49, 153
Hardy, M. A. 1993, Regression with Dummy Variables (Newbury Park, CA: SAGE)
Hayes, L. A., Gallagher, P. T., McCauley, J., et al. 2016, JGR, 122, 9841
Jain, A. K., Murty, N. M., & Flynn, P. J. 1999, ACM Comput. Surv., 31, 264
Kontar, E. P., Brown, J. C., Emslie, A. G., et al. 2011, SSRv, 159, 301
Li, R., Wang, H. N., He, H., Cui, Y. M., & Du, Z. L. 2007, ChJAA, 7, 441
McIntosh, P. S. 1990, SoPh, 125, 251
Muranushi, T., Shibayama, T., Muranushi, Y., et al. 2015, SpWea, 13, 778
Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, ApJ, 835, 156
Olden, J. D., Joy, M. K., & Death, R. G. 2004, Ecol. Model., 178, 389
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Natur, 323, 533
Stone, M. 1974, J. R. Stat. Soc., Ser. B, 36, 111
Tibshirani, R. 1996, J. R. Stat. Soc., Ser. B, 58, 267
Wainberg, M., Alipanahi, B., & Frey, B. J. 2016, J. Mach. Learn. Res., 17, 1
Wheatland, M. S. 2004, ApJ, 609, 1134
Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. 2009, Bioinformatics, 25, 714
Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, SoPh, 255, 91
Yuan, Y., Shih, F. Y., Jing, J., & Wang, H. M. 2010, RAA, 10, 785