RAA 2017 Vol. **17** No. **4**, 34 (12pp) doi: 10.1088/1674–4527/17/4/34
© 2017 National Astronomical Observatories, CAS and IOP Publishing Ltd.
http://www.raa-journal.org *http://iopscience.iop.org/raa*

Research in Astronomy and Astrophysics

Short-term solar flare prediction using multi-model integration method

Jin-Fu Liu, Fei Li, Jie Wan and Da-Ren Yu[†]

School of Energy Science and Engineering, Harbin Institute of Technology, Harbin 150001, China;
yudaren@hit.edu.cn

Received 2016 August 23; accepted 2016 November 25

Abstract A multi-model integration method is proposed to develop a multi-source and heterogeneous model for short-term solar flare prediction. Different prediction models are constructed on the basis of extracted predictors from a pool of observation databases. The outputs of the base models are normalized first because these established models extract predictors from many data resources using different prediction methods. Then weighted integration of the base models is used to develop a multi-model integrated model (MIM). The weight set that single models assign is optimized by a genetic algorithm. Seven base models and data from *Solar and Heliospheric Observatory*/Michelson Doppler Imager longitudinal magnetograms are used to construct the MIM, and then its performance is evaluated by cross validation. Experimental results showed that the MIM outperforms any individual model in nearly every data group, and the richer the diversity of the base models, the better the performance of the MIM. Thus, integrating more diversified models, such as an expert system, a statistical model and a physical model, will greatly improve the performance of the MIM.

Key words: methods: statistical — Sun: activity — Sun: magnetic fields — Sun: photosphere — Sun: flares

1 INTRODUCTION

There is a great demand for accurate flare prediction because of the effect of large solar flares on local “space weather.” Much effort has been devoted to improving short-term flare predictions.

One aspect of the work is to find more informative predictors. McIntosh (1990) proposed the McIntosh classification of sunspots to reflect the morphological characteristics of active regions and developed an expert system called Theophrastus to predict solar flares. This classification scheme and system were widely applied to later research. Bornmann & Shaw (1994) used multiple linear regression analyses to decrease the original McIntosh classification parameters from 17 to 10 while still ensuring accuracy of the observed flare rates. Gallagher et al. (2002) evaluated prediction rate using McIntosh classification, and then forecasted the occurrence of daily flares by assuming a Poisson distribution for the waiting time of X-ray flares. Cui et al. (2006) presented three kinds of predictors to describe the nonpotentiality and complexity of the photospheric magnetic field: the maximum horizontal gradient, the length of the neutral line and the number of singular points. Yu et al. (2010a) implemented a model that uses multiple-resolution predictors resulting from the decomposition of a sequence of predictors into four frequency bands using a maximal overlap discrete wavelet transform. This model reflects the trend and the changing rate of the emerging flux regions. Huang et al. (2010) constructed a C4.5 decision tree model based on predictor teams that are extracted from a dataset using rough set theory. The predictor teams used in the ensemble model not only efficiently reduce redundancy but also heighten the profitability of the information. Huang et al. (2013) constructed a metric to depict the positional relationship between active regions and predicted active longitudes, which enhance the performance of solar flare prediction. Volobuev et al. (2016) proposed that a generalized Laplacian could help predict strong flares and found that the maximum Laplacian is located near the active region polarity inversion line.

Another aspect of flare prediction is to construct more powerful models. Bradshaw et al. (1989) con-

[†] Corresponding author.

structed a three-layer back-propagation neural network named TheoNet to forecast flares. Wang et al. (2008) proposed new measurements based on solar magnetic field observations that provide more information than what can be provided by measurements based on sunspot group classification and then set up a solar flare forecasting model supported by an artificial neural network. Wheatland (2004) proposed a Bayesian approach that refines the prediction of an occurrence of a large flare during a subsequent period by using the historical record of flares within an active region together with phenomenological rules on flare statistics. Yu et al. (2010b) presented an innovative Bayesian approach to flare prediction supported by feature extraction that compared two prediction models using raw sequential data and feature-extracted data, respectively, and obtained a more comprehensive method. Li et al. (2007) presented a flare prediction model based on a support vector machine (SVM) combined with the k-nearest neighbors method. Qahwaji & Colak (2007) suggested a hybrid system that combines SVM and cascade-correlation neural networks based on the McIntosh classification so as to convert every relevant extracted sunspot to a numerical format. Colak & Qahwaji (2009) presented an automated hybrid computer platform (ASAP) for short-term prediction of significant solar flares using *Solar and Heliospheric Observatory* (*SOHO*)/Michelson Doppler Imager (MDI) images. Huang et al. (2010) proposed a flare prediction model based on sequential data by using a sliding window method to build the dynamic characteristics of the prediction model and then proposed multiscale predictors of the photospheric magnetic field. Li & Zhu (2013) considered the evolution of solar active regions and used sequential sunspot data to predict solar flares. Ahmed et al. (2013) applied machine learning and feature-selection algorithms to a set of magnetic feature properties to determine the ability to predict solar flares and the relationship between these magnetic properties and the occurrence of flares. Bloomfield et al. (2012) used X-ray flares measured by the *Geostationary Operational Environmental Satellite* (*GOES*) and McIntosh group classifications to determine the Poisson probabilities for different flare magnitudes. Muranushi et al. (2015) developed the UFCORIN platform for studying and automating the prediction of space weather, including solar flares. Shin et al. (2016) focused on the flux of strong flares and proposed models to forecast the daily maximum flux of strong flares (M- and X-class) using multiple linear regression and artificial neural network methods.

Many methods and theories have been used to predict solar flares; however, the physical mechanism of flare eruption is so complex that no one model can extract enough information and physical features from observations to use for flare prediction. Only one or a limited number of terms can be considered in the framework of a single model and then lead to good accuracy in prediction. However, if existing models could be integrated, the resulting model would include different terms with respect to the mechanism of flare eruption and different observed data. Thus, this hybrid model not only would be better at predicting solar flares but also, more importantly, would have greater adaptability and generalizing performance. This paper presents a multi-model integrated model (MIM) based on a global optimal searching method that takes advantage of individual models, including the physical model, the expert system model, machine learning model, statistical model, etc. Seven classifications were chosen to train and analyze the observed data. The MIM combines the outputs of different classifiers using a genetic algorithm such that the final output is a weighted average of different methods which emphasize different aspects they consider.

This paper is structured as follows: Experimental data are briefly explained in Section 2. In Section 3, implementation of the multi-model integration method and the construction of the model are introduced. In Section 4, base models are selected, experimental processes are explained and results are presented. Section 5 compares and analyzes the results. Finally, Section 6 presents conclusions and discusses further research.

2 DATA

The strength of a solar flare is defined in terms of five levels: A, B, C, M and X, and the influence of flares is determined by them. The total importance of a flare within a certain interval is conventionally a linear sum of those five levels

$$I_{\text{tot}} = \sum C + 10 \times \sum M + 100 \times \sum X. \qquad (1)$$

Equation (1) considers the influence of all the flares within the forward-looking period. For example, if an active region produces $C1.2$, $C2.3$, $M4.1$ and $X1.2$ flares within 48 hours, the $I_{\text{tot}} = (1.2+2.3)+10\times4.1+100\times1.2 = 164.5$ (Wang et al. 2008). Generally, a prediction model considers the eruptions of flares with significance above a threshold as a “flaring” class. Here the threshold is 10. It means that the definition of “flaring” versus “non-flaring” is the total importance above $M1.0$.

The predictors that are used in this experiment are the maximum horizontal gradient ($|\nabla_h B_z|_{\text{m}}$), the length of neutral line (L) and the number of singular points (η). They are extracted from active regions in *SOHO*/MDI full disk longitudinal magnetograms with a pixel size of $2''$ and noise level of 20 G. Active regions are selected using the following two criteria:

(1) At least one X-ray flare whose magnitude ≥C1.0 is produced in these active regions.
(2) The locations of active regions are within 30° of the solar disk center.

In general, large flares receive more attention so the first criterion aims to focus on active regions above a certain threshold. The second criterion reduces the impact of projection effects. The active regions are manually extracted from a rectangular patch on the magnetogram. The locations of the active regions associated with solar flare events are obtained from Solar Geophysical Data solar event reports (*http://www.solarmonitor.org/index.php*). A rectangular patch is used to select the active region. When two active regions are in the same patch, they are considered as one active region. Data are collected from 1996 April 15 to 2008 April 2 and the time interval of data sampling is 96 minutes from successive magnetograms.

Figure 1 shows a *SOHO*/MDI magnetogram (2011 January 11). There is a statistical relationship between observed data from photospheric magnetic field and solar flare productivity which is called a priori information in the field of probability and machine learning methods. Generalization performance could be improved by using them in a prediction model. Cui et al. (2006) pointed out three predictors (the maximum horizontal gradient, the length of neutral line, the number of singular points) and productivity of a solar flare can be fitted by a sigmoid relationship

$$Y = A_2 + \frac{A_1 - A_2}{1 + \exp\left[(X - X_0)/W\right]}\,, \qquad (2)$$

where Y is the flare productivity defined by the ratio of the number of flaring samples to the number of total samples, and X is the value of the predictor. A_1, A_2, X_0 and W are estimated from the curve-fitting process. Their values, which are shown in Table 1, are those when the threshold is M1.0. The data are preprocessed using the sigmoid function to set up a relatively simple model.

In this study, we divided the dataset into ten groups based on different active regions. Data should not be trained and tested at the same time if they are from the same active region because such data have similar statistical properties and physical features. Therefore, the use of groups of data from different active regions for cross validation is practical and the results are more credible.

3 MULTI-MODEL INTEGRATED MODEL

The MIM was constructed by training different base models and then combining the output of each model in a particular way to yield the final result. The learning strategy of a base model is a search for optimization, so the solution space of the MIM is reduced, which helps approach the best results. An MIM can be constructed using observed data of different physical mechanisms, different predictors, or various models only if a better-adapted technique is found that can combine the individual models, refine the effective information and optimize the results. The logical relationship of the model is "data and predictors-based models combination final result." Figure 2 shows the structure of the model.

3.1 Construction of MIM

The MIM has been validated as a useful application in solar flare prediction because of the diversity of single models. It can be well explained by two aspects. Firstly, on the side of machine learning and data mining techniques, structures of data have great influences on the performance of different learning methods. For example, a decision tree learning model can form a number of "IF-THEN" rules like a tree after training data so it seems to be a practical model with clear logics and regulations if the rules of data are clear as well. However, a neural network model would be better compared with a decision tree when there are messy, massive and irregular data that need to be trained. A neural network is a typical "black box" model which can adaptively set up a huge model with unreadable black rules by itself. It is convenient and easy to train this kind of data, not a complex and slow process of model construction by decision tree, even overfitting. As another example, assume that attributes of data are mutually independent and a naive Bayes model can perform perfectly, otherwise, it will be far less desirable than the performance of the Bayesian network. Secondly, on the side of physical observation of a solar flare, relationships between solar flare productivity and observed data are diverse as well as predictors estimated from the same observation. For example, the maximum horizontal gradient ($|\nabla_h B_z|_{\rm m}$), the length of neutral line (L) and the number of singular points (η) are extracted from active regions in *SOHO*/MDI full disk longitudinal magnetograms while the magnetic shear ($L_{\rm s}$), the current ($I_{\rm tot}$) and the current helicity ($H_{\rm tot}$) are from active regions in vector magnetograms. McIntosh sunspot classification divides the shapes of sunspots into seven categories and extracts McIntosh parameters, which act as proxies for the magnetic properties. It is reasonable to

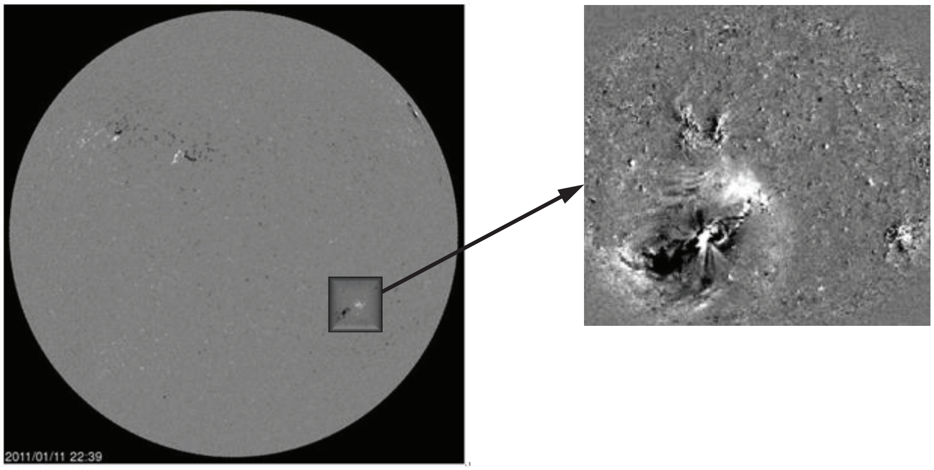

Fig. 1 (*left*) *SOHO*/MDI full-disk magnetogram obtained on 2011 January 11; (*right*) an active region on the magnetogram.

Table 1 Values of Parameters in Sigmoid Functions

Threshold	Forward-looking period	Predictor	A_1	A_2	X_0	W
$I_{\rm tot} = 10$	48 (h)	$\vert\nabla_h B_z\vert_{\rm m}$	0.164	0.738	0.360	0.066
		L	0.062	0.848	763.08	382.97
		η	–0.196	0.730	9.343	22.663

analyze the mechanism of flare eruption from a holistic perspective and apply it to a prediction model and it is available for multiple models.

The MIM combines base models via the linear sum of their outputs. The weights of these models are optimized by a genetic algorithm, and the fitness function is changed with the optimization index to obtain a good result.

3.2 Weighted Integration of Multiple Models Based on a Genetic Algorithm

The ultimate purpose of the MIM is to enhance the strength and eliminate the weakness of the base models that it comprises, so the technique used to combine them and the guarantee of desirable results are significant in the performance of the MIM. We used a weighted mean of the output of the individual models and a genetic algorithm in combining them, and optimized the final output. This proved to be an efficient approach to combining the base models.

Due to different data resources and predictors generated together with various modeling constructions, the effectiveness of based models for flare prediction may not be completely the same. Thus the output normalization of each selected base model before weight assignment is necessary which ensures implementation of integration. After base models have been built, normalized outputs of base models for one sample x can be described as

$$\begin{aligned} &h_1(x), h_2(x), \ldots, h_T(x); \\ &h_i(x) \in \{-1, 1\}, \quad i = 1, 2, \ldots, T. \end{aligned} \tag{3}$$

Each model i labels x as “flaring” or “non-flaring,” as explained above, where 1 means a sample is assigned to “flaring” class and -1 means a sample belongs to “non-flaring” class.

A standard and easy-to-realize scheme of base model integration is needed after base model normalization because of the heterogeneity of base models and complexity of the interconnection among them. Weighted integration for base models by using a genetic algorithm can be easily applied in the process because the relationship of base models for flare prediction can be described by weight assignment and the genetic algorithm acts to optimize weights for yielding a better prediction result of MIM.

Genetic algorithm search procedures are loosely based on the principal of natural selection: they “evolve” good feature subsets by using random perturbations of a current list of candidate subsets. Although different implementations of genetic algorithms vary in detail, they typically share the following structure: The algorithm operates by iteratively updating a pool of hypothesized solutions, called the population. At each iteration, all mem-

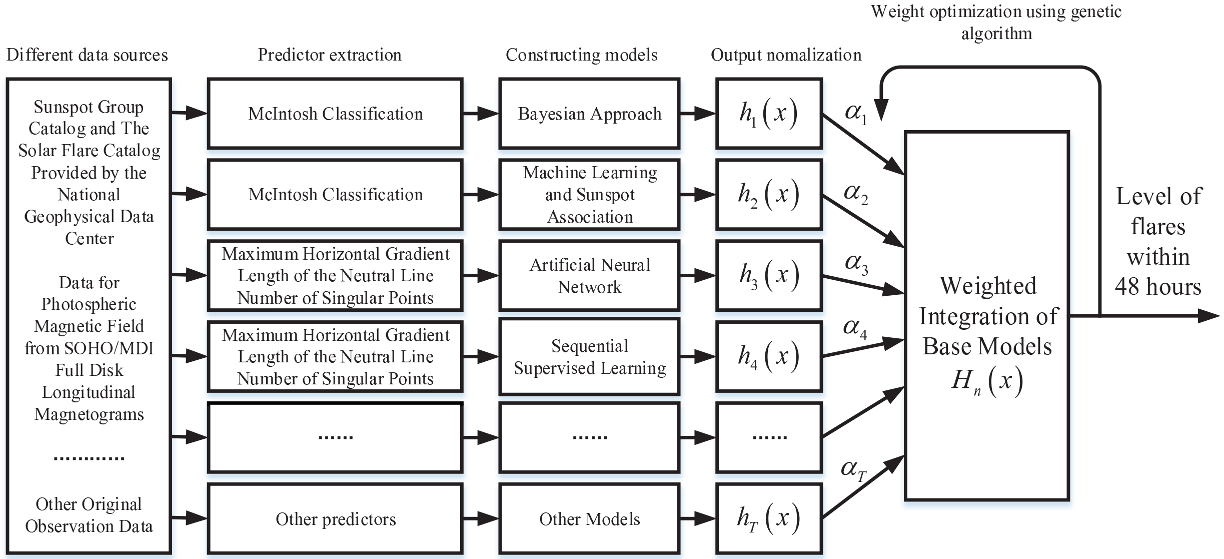
Weight optimization using genetic algorithm
Different data sources
Predictor extraction
Constructing models
Output nomalization
Sunspot Group Catalog and The Solar Flare Catalog Provided by the National Geophysical Data Center
Data for Photospheric Magnetic Field from SOHO/MDI Full Disk Longitudinal Magnetograms
....
Other Original Observation Data
McIntosh Classification
McIntosh Classification
Maximum Horizontal Gradient Length of the Neutral Line Number of Singular Points
Maximum Horizontal Gradient Length of the Neutral Line Number of Singular Points
......
Other predictors
Bayesian Approach
Machine Learning and Sunspot Association
Artificial Neural Network
Sequential Supervised Learning
......
Other Models
$h_1(x)$
$h_2(x)$
$h_3(x)$
$h_4(x)$
......
$h_T(x)$
α_1
α_2
α_3
α_4
α_T
Weighted Integration of Base Models $H_n(x)$
Level of flares within 48 hours

Fig. 2 Schematic diagram of the MIM.

bers of the population are evaluated according to the fitness function. A new population is then generated by probabilistically selecting the most fit individual from the current population. Some of these selected individuals are carried forward into the next generation population intact. Others are used as the basis for creating new offspring individuals by applying genetic operations such as crossover and mutation. Figure 3 illustrates a diagram of how a genetic algorithm operates.

In this case, evolution of the population updates weights for a better result of the MIM but the outputs of base models will not change when the weights change. Thus, initializing the population in terms of encoding only depends on weights.

Using double encoding (double precision floating point numbers in a population) to initialize a population (chrom) is described as

$$W_{\mathrm{Chrom}} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1T} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2T} \\ \dots & \dots & \dots & \dots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mT} \end{bmatrix}, \qquad (4)$$
$$\sum_{i=1}^{\mathrm{T}} \alpha_{ki} = 1, \qquad k = 1, 2, \dots, m,$$

where m stands for the number of individuals in the population and each gene $i = 1, 2, \dots, T$ in an individual stands for the weight of the single model.

Then in the process of constructing fitness function $f = \mathrm{Obj}(h, W_{\mathrm{Chrom}})$ and evaluating individuals, the fitness value is described as

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \dots \\ f_{\mathrm{m}} \end{bmatrix}. \qquad (5)$$

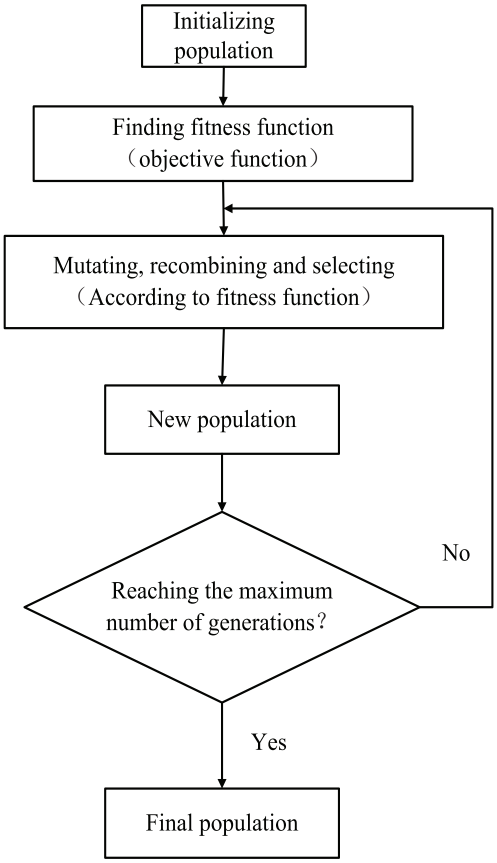
Initializing population
Finding fitness function （objective function）
Mutating, recombining and selecting （According to fitness function）
New population
Reaching the maximum number of generations?
No
Yes
Final population

Fig. 3 Diagram illustrating how a genetic algorithm operates.

The fitness function is constructed as follows:

The sample is assigned to the class in the MIM which is based on the weighted mean of single models and the subscript T is also the number of models we selected

$$H(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^{\mathrm{T}} \alpha_i h_i(x) > 0, i = 1, 2, \dots, T; \\ -1, & \text{otherwise}. \end{cases} \tag{6}$$

Here 1 means a sample is assigned to "flaring" class and -1 means a sample belongs to "non-flaring" class, the same as above. The element in $(\alpha_1, \alpha_2, \dots, \alpha_T)$ is the weight of each model.

Testing results for each sample (x_i, y_i) by those base models $h_1(x), h_2(x), \dots, h_T(x)$ compose a matrix

$$h = \begin{bmatrix} h_{11}(x) & h_{12}(x) & \dots & h_{1T}(x) \\ h_{21}(x) & h_{22}(x) & \dots & h_{2T}(x) \\ \dots & \dots & \dots & \dots \\ h_{n1}(x) & h_{n2}(x) & \dots & h_{nT}(x) \end{bmatrix}, \tag{7}$$

where n is the number of total samples. The final output of the MIM can be described as

$$\begin{aligned} H &= W_{\text{Chrom}} \times h^{\mathrm{T}} \\ &= \begin{bmatrix} H_1(x_1) & H_1(x_2) & \dots & H_1(x_n) \\ H_2(x_1) & H_2(x_2) & \dots & H_2(x_n) \\ \dots & \dots & \dots & \dots \\ H_{\mathrm{m}}(x_1) & H_{\mathrm{m}}(x_2) & \dots & H_{\mathrm{m}}(x_n) \end{bmatrix}. \end{aligned} \tag{8}$$

As expatiated above, each individual has a classification for each data sample, so the fitness number is estimated by H according to performance evaluation such as area under curve (AUC) and Heidke Skill Score (HSS), which will be introduced in Section 4.

After determining the initial population (chrom) and fitness function, selection, mutation and recombination are generally taken into consideration to optimize the population to achieve optimal performance. A stochastic universal sampling method is used to select an individual with high fitness value and individuals mutate with a uniform random probability while two matched individuals are recombined by the two-point crossover method.

A stochastic universal sampling method is an update of the roulette rule which can select all the individuals of the next generation in an iteration. Points are evenly distributed on a representation of a roulette wheel, the number of which is equal to the population size. Uniform mutation is defined such that an original encoded gene of an individual is replaced by a random value from a certain range, which then turns into a new genetic code. Two-point crossover refers to two intersecting points selected in a couple of matched individuals and encoded data between them being subsequently exchanged with each other. Assume that a couple is A and B:

$$A : (\alpha_{A1}, \alpha_{A2}, \dots, \alpha_{AT}); B : (\alpha_{B1}, \alpha_{B2}, \dots, \alpha_{BT})$$

There are $(T-1)$ possible positions for cross points which are between two adjacent codes. The first point is between α_2 and α_3, and another is between α_5 and α_6. After recombining, two individuals are turned into:

$$\begin{aligned} A' &: (\alpha_{A1}, \alpha_{A2}, \alpha_{B3}, \alpha_{B4}, \alpha_{B5}, \alpha_{A6}, \dots, \alpha_{AT}); \\ B' &: (\alpha_{B1}, \alpha_{B2}, \alpha_{A3}, \alpha_{A4}, \alpha_{A5}, \alpha_{B6}, \dots, \alpha_{BT}). \end{aligned}$$

New population (Chrom) W^2_{Chrom} is reborn when selection, mutation and recombination are completed in the first generation. Result T^2 can be estimated based on W^2_{Chrom} and then the second generation ends and W^3_{Chrom} is generated. This cycle repeats before approaching the maximum number of generations and T^g is the final result of the MIM.

4 EXPERIMENT AND RESULTS ANALYSIS

4.1 Performance Evaluation

Three predictors, observed data of which are the maximum horizontal gradient ($|\nabla_h B_z|_{\mathrm{m}}$), the length of neutral line (L) and the number of singular points (η), are turned into to solar flare productivity through the sigmoid relationship in Equation (2). A solar flare prediction can be treated as a binary classification task, therefore, after training each sample there are four different possible outcomes shown in Table 2. We consider flaring samples as the Positive class and non-flaring samples as the Negative class. Samples correctly classified as "Positive" are defined as "True Positive" (TP) and those incorrectly classified as "Positive" are defined as "False Positive" (FP). In turn, samples correctly predicted as "Negative" are called "True Negative" (TN) while samples wrongly predicted as "Negative" are called "False Negative" (FN) (Witten & Frank 2005).

Based on the confusion matrix explained above, there are definitions of TP rate and TN rate which measure the accuracy of prediction. TP rate is the ratio of the number of samples which are correctly classified as Positive to the number of samples that belong to the actual Positive class

$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{9}$$

TN rate is the ratio of the number of samples that are correctly classified as Negative to the number of samples that belong to the actual Negative class

$$\text{TN rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{10}$$

Table 2 Different Outcomes of Two-class Prediction

	Predicted positive class	Predicted negative class
Actual positive class	True Positive	False Negative
Actual negative class	False Positive	True Negative

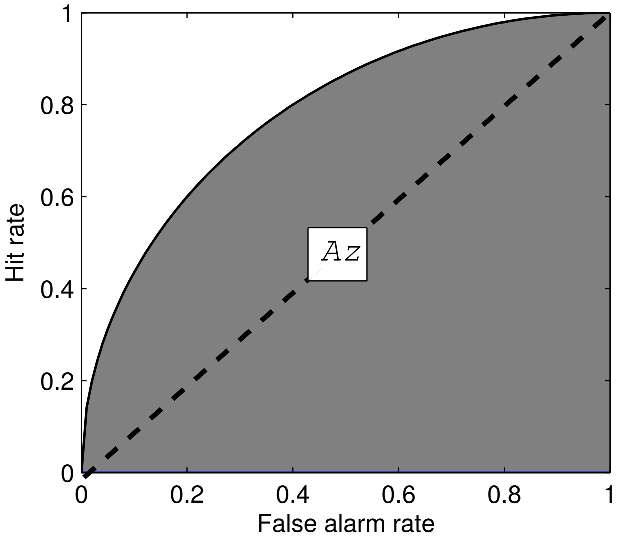

Fig. 4 Schematic diagram of AUC. Surface of the area is ROC and the area of the shadow is AUC.

Imbalanced data can be measured by AUC which is area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of the FPrate on the X axis versus the TPrate on the Y axis. It shows that the differences between the FPrate and the TPrate are based on different rules. HSS is usually used to quantify the performance which can be defined with

$$\mathrm{HSS} = \frac{\mathrm{PC} - \mathrm{E}}{1 - \mathrm{E}}. \tag{11}$$

$$\mathrm{PC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{N}},$$
$$\mathrm{E} = \frac{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP}) + (\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}{\mathrm{N}^2}, \tag{12}$$

where PC is accuracy of prediction and E is accuracy of a random forecast (Jolliffe & Stephenson 2012).

4.2 Base Model Selection

Attempts have been made to combine physical models, expert system models and statistical models with machine learning models, but more advanced techniques are needed. We explored combining various machine learning models to show that multiple models provide a distinctly better performance. Two characteristics that determine the outcome of the MIM are the accuracy and the variety of base models. The seven base models of the MIM were based on naive Bayes, SVM, Bayes network, C4.5 decision tree, radial basis function network, multilayer perceptron method and sequential minimal optimization method.

A Bayesian network together with the naive Bayes algorithm forms a good base on which to build probabilistic reasoning models. Bayesian learning algorithms that can be used to calculate explicit probabilities for hypotheses are among the most practical approaches to certain types of learning problems. Bayesian inference is based on the assumption that attributes are conditionally independent and it uses Bayes theorem. In contrast to the naive Bayes algorithm, a Bayesian network allows the assumption of conditional independence to be applied to a subset of variables in the naive Bayes algorithm. A directed acyclic graph depicts the relationships between the attributes and simplifies the evaluation of the joint probability density function. When the conditional probability table for attributes is explicit, the setup of the network is complete.

SVM is an algorithm that emphasizes structural risk minimization theory. An SVM can operate like a linear model to obtain the description of a nonlinear boundary of a dataset using a nonlinear mapping transform. A kernel function converts input in low-dimensional space to output in higher-dimensional space. A classifier that uses a sequential minimal optimization method can dramatically improve the performance of quadratic programming in an SVM model. The LibSVM classifier is another technique for obtaining SVM classification.

The perceptron learning rule is built by a hyperplane alone, with a group of weights assigned to each attribute, including an extended attribute equal to 1. Data are classified into one class when the sum of the weights of an attribute is a positive number and into another class when the sum is a negative number. The attributes are reweighted if the samples are classified incorrectly until the classification is correct. However, nonlinear separated data will result in nonconvergence of the classification function. A multilayer perceptron classifier can handle data that are linearly inseparable by constructing a network of perceptron classifiers that defines the nonlinear boundary of the dataset. A radial basis function network is another kind of feed forward network that uses a Gaussian function as its activation function and a sigmoid function to transform the classification.

The C4.5 decision tree is an extension of the ID3 algorithm. It determines the affiliations of the nodes by using an information gain ratio. Use of the post-pruning rule improves the generation performance of the C4.5 model. C4.5 can discretize continuous attributes while the ID3 decision tree is restricted to discrete attribute processing. C4.5 can also handle training examples with missing attribute values.

4.3 Performance of Multi-Model Integrated Model

We use ten-fold cross validation (Kohavi et al. 1995) to validate performance of single models and the MIM. The dataset is decomposed into ten folds based on different active regions, nine folds are trained by models and one fold is applied to test them in turn until each fold has been used as test data once in cross validation. Ten groups of prediction outcomes are recorded for the individual models and the MIM.

The performances of individual models and the MIM as determined by different evaluation methods, i.e., TPrate, TNrate, AUC and HSS, are compared in Tables 3–6. The results of the ten groups and their means are given. The performance of the MIM in eight and seven groups was the best for TPrate and TNrate evaluations, respectively. Likewise, the performance of the MIM in seven and five groups was the best for AUC and HSS evaluations, respectively.

5 ANALYSIS OF EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Influence of Base Model Diversity on Performance of MIM

An MIM, acting as a combination of single models, will outperform individual models through theoretic analysis below. Firstly, physical phenomena related to a solar flare have been described and translated into different values of several predictors. Each model shows great power in classification of a flare. Some of them were used in construction of flare prediction before and proved to be desirable techniques. Secondly, individual models using different methods reflect diversity in training data and making rules. It means not only the structure of the models is different, but various results of prediction they support as well. More importantly, difference in results is of great importance in combination and integrated decision while the same results are insignificant in multiple models no matter if they are true or false.

AUC and HSS measure the accuracy of the base models while entropy E, as a correlation metric, measures their diversity. A high correlation among the individual models shows they lack diversity and a low correlation shows they are very diverse. If the models are unpaired, entropy (Kuncheva 2004) is a measurement that can be applied to individual models

$$\mathrm{Div} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{L-\lfloor L/2\rfloor}\min\left\{\sum_{i=1}^{L}y^{j,i}, L-\sum_{i=1}^{L}y^{j,i}\right\} \tag{13}$$

where N is the number of total samples and L is the number of single models. $\lfloor a\rfloor$ is the biggest integer that is less than a and $y^{j,i}$ are the oracle outputs by individual models h_i for every sample $j = 1, 2...N$. It seems that if the predictions of the sample are the same, the returned value of Div will be 0, so the larger the value of E is, the more diverse single models are.

The diversity of the base models for which the data were divided into ten active region groups is presented in Table 7.

To demonstrate the above conclusion, Figure 5 shows the diversity of the base models and the optimization gain of the MIM over that of the base models. The top panel shows the diversity of the base models for the ten data groups and the middle and bottom panels show the average AUC and HSS optimization gain with the MIM versus that of seven base models for the same ten data groups, respectively. The average AUC and HSS gains with the MIM are defined as follows

$$\begin{aligned} \mathrm{gain}_{\mathrm{AUC}} &= \mathrm{AUC}_{\mathrm{MIM}} - \overline{\mathrm{AUC}_{\mathrm{base\ models}}}, \\ \mathrm{gain}_{\mathrm{HSS}} &= \mathrm{HSS}_{\mathrm{MIM}} - \overline{\mathrm{HSS}_{\mathrm{base\ models}}}. \end{aligned} \tag{14}$$

The linear correlation between diversity and the MIM gain of AUC and HSS is shown in Figure 6.

There are three main reasons that can explain better results of the MIM. Firstly, the base models have good performance in testing, which all have an accuracy above 0.6. Secondly, the diversity of models is good. Several algorithmic methods are used in constructing prediction models and the generalization performance of MIM is far superior to that of single models.

In Figure 5, the diversity and average gain for AUC and HSS have similar behavior and in Figure 6 there is a significant linear correlation between diversity and optimization gain of AUC and HSS, i.e., greater diversity leads to higher gain and less diversity leads to lower gain. Thirdly, the genetic algorithm which is used in optimization of weights is able to search for the global optimal solution without making rules during processing; therefore, it can automatically find a group of weights for each

Table 3 Performance of Individual Models and MIM Using TPrate

	RBFN	SVM	C4.5	SMO	BayesNet	NaiveBayes	Multi-perceptron	MIM
1	0.6379	0.6538	0.6305	0.6559	0.629	0.6357	0.6284	0.6612
2	0.6851	0.6953	0.6953	0.7221	0.6908	0.6895	0.7297	0.7297
3	0.5839	0.569	0.5749	0.5764	0.5968	0.579	0.5619	0.5928
4	0.5751	0.5904	0.5941	0.6094	0.5876	0.5826	0.6068	0.6155
5	0.626	0.617	0.6212	0.5861	0.6313	0.6047	0.6091	0.6317
6	0.5835	0.5529	0.5544	0.5578	0.596	0.5762	0.5763	0.5925
7	0.6657	0.6784	0.649	0.6585	0.6601	0.6372	0.6647	0.6809
8	0.5886	0.6251	0.6462	0.6092	0.6591	0.6196	0.6374	0.652
9	0.6202	0.6578	0.6508	0.6235	0.6814	0.6387	0.6407	0.6814
10	0.5929	0.6383	0.6389	0.6185	0.6553	0.6349	0.6428	0.6574
Mean	0.61589	0.6278	0.62553	0.62174	0.63874	0.61981	0.62978	0.64951

Table 4 Performance of Individual Models and MIM Using TNrate

	RBFN	SVM	C4.5	SMO	BayesNet	NaiveBayes	Multi-perceptron	MIM
1	0.8213	0.8372	0.8139	0.8393	0.8124	0.8191	0.8118	0.8446
2	0.8685	0.8787	0.8787	0.9055	0.8742	0.8729	0.9131	0.9131
3	0.7673	0.7524	0.7583	0.7598	0.7802	0.7624	0.7453	0.7762
4	0.7585	0.7738	0.7775	0.7928	0.771	0.766	0.7902	0.7989
5	0.8094	0.8004	0.8046	0.7695	0.8147	0.7881	0.7925	0.8151
6	0.7669	0.7363	0.7378	0.7412	0.7794	0.7596	0.7597	0.7759
7	0.8491	0.8618	0.8324	0.8419	0.8435	0.8206	0.8481	0.8643
8	0.772	0.8085	0.8296	0.7926	0.8425	0.803	0.8208	0.8354
9	0.8036	0.8412	0.8342	0.8069	0.8648	0.8221	0.8241	0.8648
10	0.7763	0.8217	0.8223	0.8019	0.8387	0.8183	0.8262	0.8408
Mean	0.7993	0.8112	0.8089	0.8051	0.8221	0.8032	0.8132	0.8323

Table 5 Performance of Individual Models and MIM Using AUC

	RBFN	SVM	C4.5	SMO	BayesNet	NaiveBayes	Multi-perceptron	MIM
1	0.7296	0.7455	0.7222	0.7476	0.7207	0.7274	0.7201	0.7529
2	0.7768	0.787	0.787	0.8138	0.7825	0.7812	0.8214	0.8214
3	0.6756	0.6607	0.6666	0.6681	0.6885	0.6707	0.6536	0.6845
4	0.6668	0.6821	0.6858	0.7011	0.6793	0.6743	0.6985	0.7072
5	0.7177	0.7087	0.7129	0.6778	0.723	0.6964	0.7008	0.7234
6	0.6752	0.6446	0.6461	0.6495	0.6877	0.6679	0.668	0.6842
7	0.7574	0.7701	0.7407	0.7502	0.7518	0.7289	0.7564	0.7726
8	0.6803	0.7168	0.7379	0.7009	0.7508	0.7113	0.7291	0.7437
9	0.7119	0.7495	0.7425	0.7152	0.7731	0.7304	0.7324	0.7731
10	0.6846	0.73	0.7306	0.7102	0.747	0.7266	0.7345	0.7491
Mean	0.70759	0.7195	0.71723	0.71344	0.73044	0.71151	0.72148	0.74091

base model and approach the best-matched fusion strategy.

5.2 Performance of Other Multiclass Models

Other typical ensemble classification methods, such as the AdaBoost and MultiBoost algorithms, can be applied to prediction models. The AdaBoost algorithm is used to construct a "strong" classifier by linearly combining "simple" and "weak" classifiers using reweighting. It often dramatically improves the performance of weak classifiers but it sometimes overfits. AdaBoost optimizes the solution with each iteration, in which the weights of trained samples misclassified by a weak classifier are increased and the weights of samples correctly classified are decreased. The next classifier uses the new data dis-

Table 6 Performance of Individual Models and MIM Using HSS

	RBFN	SVM	C4.5	SMO	BayesNet	NaiveBayes	Multi-perceptron	MIM
1	0.3191	0.3621	0.2962	0.3684	0.3123	0.3651	0.4087	0.3925
2	0.4906	0.5154	0.4874	0.554	0.4706	0.4985	0.5647	0.5537
3	0.2667	0.2588	0.2695	0.2651	0.2813	0.2751	0.2699	0.2976
4	0.2453	0.2841	0.2775	0.3178	0.2607	0.2782	0.2706	0.3183
5	0.3453	0.3545	0.3409	0.2944	0.3558	0.3491	0.3755	0.3786
6	0.1838	0.18	0.1698	0.18	0.2185	0.2205	0.1768	0.2258
7	0.3384	0.3764	0.3497	0.3364	0.3212	0.3269	0.2925	0.3733
8	0.205	0.3028	0.3087	0.2662	0.3257	0.2995	0.2521	0.3195
9	0.2848	0.4007	0.3587	0.3303	0.4563	0.394	0.3186	0.437
10	0.2659	0.3735	0.4056	0.3339	0.424	0.3935	0.4034	0.4246
Mean	0.29449	0.34083	0.3264	0.32465	0.34264	0.34004	0.33328	0.37209

Table 7 Diversity of Base Models

Data groups	1	2	3	4	5	6	7	8	9	10
Div	0.112	0.085	0.079	0.116	0.107	0.093	0.111	0.110	0.141	0.106

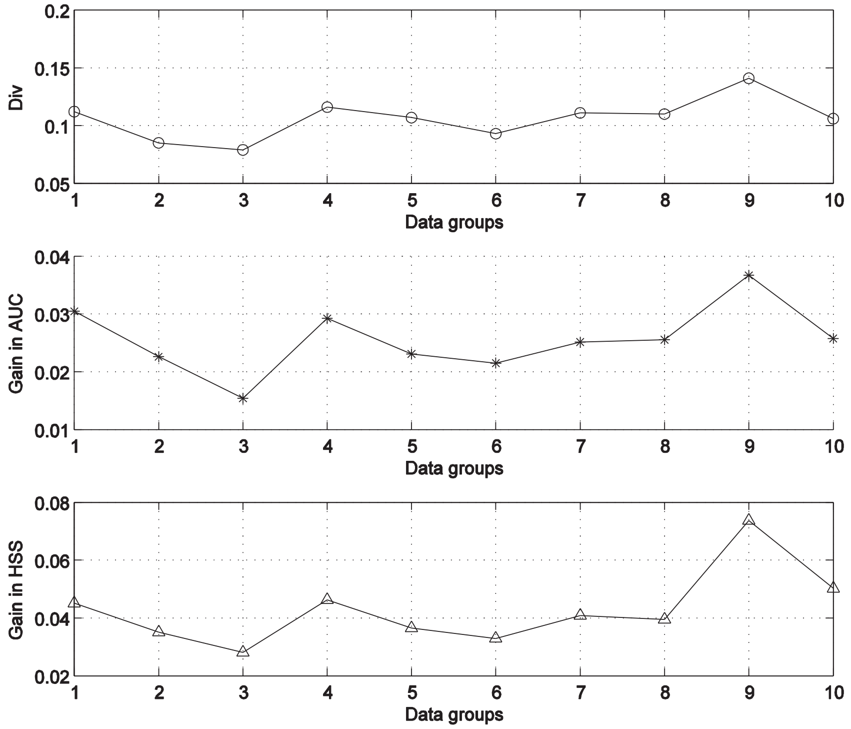

Fig. 5 Diversity of base models and average AUC and HSS gain of the MIM.

tribution and the process is repeated. MultiBoost combines AdaBoost with wagging. It is able to harness both the high bias of AdaBoost and the variance reduction by wagging.

However, the performance of MIM was better than that of AdaBoost and MultiBoost, as seen in Table 8. One reason for this result is that the AdaBoost and MultiBoost methods use the greedy method, a local searching optimization algorithm that can guarantee the best solution in every iteration but it often cannot achieve a global optimal solution. A genetic algorithm is a global searching optimization method that yields better results than other algorithms in a solar flare prediction model. Another reason that the performance of the MIM is better than

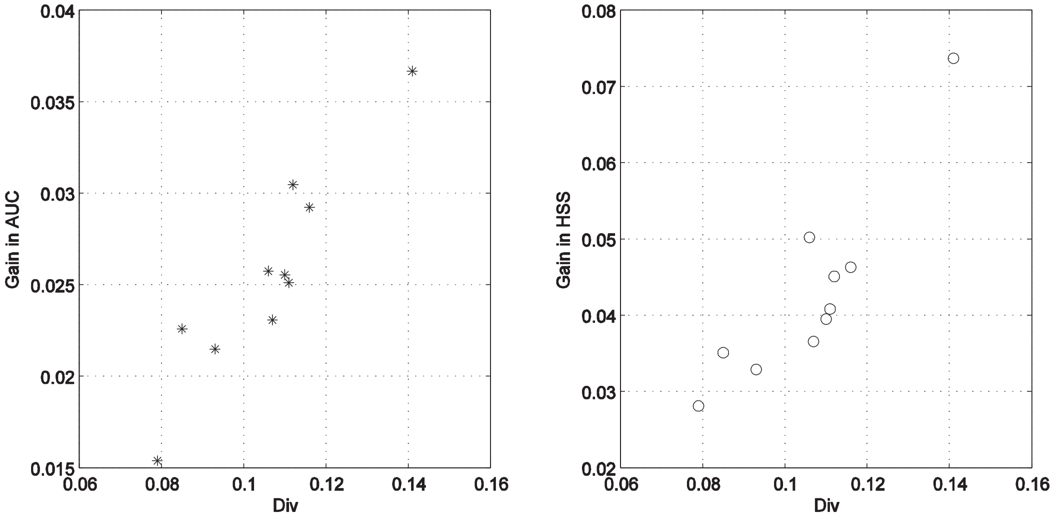

Fig. 6 Linear correlation between diversity and AUC and HSS gain.

Table 8 Performance of Different Multiclass Models

	AdaBoostM1	MultiBoostAB	Voting	MIM
AUC	0.719 ± 0.043	0.701 ± 0.054	0.725 ± 0.041	0.741 ± 0.042
HSS	0.321 ± 0.082	0.311 ± 0.103	0.342 ± 0.087	0.372 ± 0.090

AdaBoost and MultiBoost is that the latter two construct several base models using one classifier such as a decision tree. The differences among the three models are in the distribution of data and their weighting, not the classification methods. This means that the diversity of multiple models may be worse than that of models that use different classifiers. The MIM not only performs better than other multiclass models but it also is extremely versatile and extensible, and other prediction models such as an expert system or a statistical model can be integrated into it. The voting average of a multiclass model is taken into consideration and compared with the weighted average of the model after optimization. Due to the random matched fusion of a voting average, its performance will not be better than that of weighted average fusion.

6 CONCLUSIONS

It is difficult to find a model that can preserve complementary information in *SOHO*/MDI full-disk longitudinal magnetograms and also be used in classification or in the construction of readable rules to integrate several individual models into a “strong” model and obtain a good result. The MIM uses a genetic algorithm to avoid this drawback. The weights of the individual models, which are automatically created by the genetic process, act as a relationship among the models. The process of searching for an optimal solution can be viewed as integrating the best rules of the individual models without human participation. The rules set up by the models are unreadable but they are available to the MIM for it to achieve a better result than the individual models. Once the initial values, optimization direction and optimization objects are determined, the relationship among the individual models can be constructed.

To demonstrate the improvement in performance of the MIM, the models and classifiers were studied in light of data mining techniques. If the MIM comprises physical models, expert models and statistical models, their diversity would greatly improve the MIM and the MIM would perform better. Future work includes integrating more models into the hybrid system and increasing the generalization performance of the model.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant No. 11078010). We thank the *SOHO*/MDI consortium for the data. *SOHO* is a project of international cooperation between the European Space Agency (ESA) and NASA.

References

Ahmed, O. W., Qahwaji, R., Colak, T., et al. 2013, Sol. Phys., 283, 157

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJ, 747, L41

Bornmann, P. L., & Shaw, D. 1994, Sol. Phys., 150, 127

Bradshaw, G., Fozzard, R., & Ceci, L. 1989, Advances in Neural Information Processing Systems, 1, 248

Colak, T., & Qahwaji, R. 2009, Space Weather, 7, S06001

Cui, Y., Li, R., Zhang, L., He, Y., & Wang, H. 2006, Sol. Phys., 237, 45

Gallagher, P. T., Moon, Y.-J., & Wang, H. 2002, Sol. Phys., 209, 171

Huang, X., Yu, D., Hu, Q., Wang, H., & Cui, Y. 2010, Sol. Phys., 263, 175

Huang, X., Zhang, L., Wang, H., & Li, L. 2013, A&A, 549, A127

Jolliffe, I. T., & Stephenson, D. B. 2012, Forecast Verification: A Practitioner's Guide in Atmospheric Science, Second Edition

Kohavi, R., et al. 1995, A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection, in International Joint Conference on Artificial Intelligence, 14, 1137

Kuncheva, L. I. 2004, Combining Pattern Classifiers: Methods and Algorithms (John Wiley & Sons)

Li, R., Wang, H.-N., He, H., Cui, Y.-M., & Du, Z. L. 2007, ChJAA (Chin. J. Astron. Astrophys.), 7, 441

Li, R., & Zhu, J. 2013, RAA (Research in Astronomy and Astrophysics), 13, 1118

McIntosh, P. S. 1990, Sol. Phys., 125, 251

Muranushi, T., Shibayama, T., Muranushi, Y. H., et al. 2015, Space Weather, 13, 778

Qahwaji, R., & Colak, T. 2007, Sol. Phys., 241, 195

Shin, S., Lee, J.-Y., Moon, Y.-J., Chu, H., & Park, J. 2016, Sol. Phys., 291, 897

Volobuev, D., Makarenko, N., & Knyazeva, I. 2016, Generalized Laplacian for Magnetograms of Solar Active Region as Possible Predictor of Strong Flare, in Journal of Physics: Conference Series, 675, 032027

Wang, H. N., Cui, Y. M., Li, R., Zhang, L. Y., & Han, H. 2008, Advances in Space Research, 42, 1464

Wheatland, M. S. 2004, ApJ, 609, 1134

Witten, I. H., & Frank, E. 2005, Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann)

Yu, D., Huang, X., Hu, Q., et al. 2010a, ApJ, 709, 321

Yu, D., Huang, X., Wang, H., et al. 2010b, ApJ, 710, 869