Research in Astron. Astrophys. **2013** Vol. **13** No. **9**, 1118–1126
http://www.raa-journal.org *http://www.iop.org/journals/raa*

Research in Astronomy and Astrophysics

Solar flare forecasting based on sequential sunspot data *

Rong Li and Jie Zhu

School of Information Science, Beijing Wuzi University, Beijing 101149, China; *lirong@bao.ac.cn*

Received 2012 September 12; accepted 2013 April 9

Abstract It is widely believed that the evolution of solar active regions leads to solar flares. However, information about the evolution of solar active regions is not employed in most existing solar flare forecasting models. In the current work, a short-term solar flare forecasting model is proposed, in which sequential sunspot data, including three days of information about evolution from active regions, are taken as one of the basic predictors. The sunspot area, the McIntosh classification, the magnetic classification and the radio flux are extracted and converted to a numerical format that is suitable for the current forecasting model. Based on these parameters, the sliding-window method is used to form the sequential data by adding three days of information about evolution. Then, multi-layer perceptron and learning vector quantization are employed to predict the flare level within 48 h. Experimental results indicate that the performance of the proposed flare forecasting model works better than previous models.

Key words: Sun: flares — sunspots — machine learning

1 INTRODUCTION

As one kind of solar electromagnetic storm (Wang et al. 2012), a solar flare is an intense and sudden release of energy which can have a significant impact on the reliability of space-borne and ground-based technological systems. In order to protect these systems from disturbances in the space environment, probabilitic forecasting flare models are still the main tools employed in space environment services.

Up to now, a number of flare forecasting approaches and systems have been developed in which sunspot-group characteristics are considered. Based on the McIntosh (1990) classification system, an expert system was developed by the Space Weather Prediction Center of the National Oceanic and Atmospheric Administration (NOAA). Using historical averages of flare numbers for McIntosh classifications, Gallagher et al. (2002) developed a flare prediction system to evaluate the probability for each active region to produce C-, M- or X-class flares. Wheatland (2005) proposed a Bayesian approach for flare prediction, in which flaring records of an active region together with phenomenological rules of flare statistics refine an initial prediction for the occurrence of a subsequent big flare. Qahwaji & Colak (2007) proposed a short-term solar flare prediction model using a machine learning method. Wang & Zhang (1994) developed a multi-discrimination method for flare forecast with observational data on sunspots, 10 cm radio flux and longitudinal magnetic fields, and Zhu & Wang (2003) presented a verification of Wang & Zhang's method. Han et al. (2001) predicted solar

∗ Supported by the National Natural Science Foundation of China.

flares using a fuzzy clustering method, which builds a relation among several parameters describing sunspot regions. Li et al. (2007) designed a forecasting model using a support vector machine and the k nearest neighbor method, in which the sunspot area, the sunspot magnetic classification, and the McIntosh class of sunspot group and 10 cm solar radio flux were chosen. Wang et al. (2009) studied the free energy storage process in solar active regions by using a logistic model.

Previous works demonstrate that there is a clear correlation between the characteristics of sunspots and flaring occurrences. These methods are based on the current information about sunspot properties. The influence of previous parameters describing flare occurrence is not considered. However, previous information is also very important for solar flare forecasting. In solar flare forecasting models based on the characteristics of photospheric magnetic parameters, information about evolution of predictors has been added, which has improved performance (Yu et al. 2009; Li 2011). In this work, we aim to design a forecasting model that uses sequential sunspot parameters as predictors. In this system, the time series data are appended with sunspot parameters using a sliding window technique, and neural network methods are employed as forecasting methods. To estimate the performance of information about evolution, the model is built on a large-scale data set spanning all of Solar Cycle 23.

The rest of the paper is organized as follows: the data and statistical method are described in Section 2. The sliding window technique is introduced in Section 3. The neural network methods are discussed in Section 4. Experimental results are reported in Section 5. Finally, concluding remarks are given in Section 6.

2 DATA DESCRIPTION

In this work, the data are sourced from the publicly available sunspot group and the solar flare catalog. Data on solar flares are from the *GOES* satellites, which are provided by the National Geophysical Data Center (NGDC) and can be downloaded at *http://www.ngdc.noaa.gov/stp/SOLAR/ftpsolarflares.html*. The importance of flares is conventionally described with indexes of C, M or X. Considering the selected active regions producing at least one C1.0 flare, C1.0 is taken as the unit for measuring total X-ray importance. Within a certain time interval, the total importance, $I_{\rm tot}$, is defined as

$$I_{\rm tot} = 1.0 * \sum C + 10 * \sum M + 100 * \sum X \,. \tag{1}$$

Here, the threshold of $I_{\rm tot}$ is supposed to be 10, i.e. M1.0 equivalent. The forward looking period is taken to be 48 h, which is long enough for the evolution of a sunspot active region (Wang et al. 2008).

Solar sunspot data are derived from the daily active region summary report of NOAA and downloaded from *http://www.swpc.noaa.gov/ftpmenu/forecasts/SRS.html*. In this report, the solar observation of a sunspot active region is recorded, including number, location, area and classification. There are two main classification systems for sunspots: Mount Wilson and McIntosh. Mount Wilson classification is based on the distribution of magnetic polarities within spot groups, while McIntosh classification depends on the size, shape and spot density of sunspots. The data selected in our work are sunspot area, Mount Wilson classification and McIntosh classification. In addition, the f10.7 flux is included, because it has a close relation with solar activity. The flux data can be downloaded at *ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/*.

The initial values of predictors are unfit for direct inputs to forecasting models, so we calculate a statistic on the correlation of these predictors with a solar flare by calculating their flare productivities.

The flare productivities of the magnetic class and the McIntosh class of a sunspot group can be estimated by the ratio between the number of flare bursts and that of the total samples in a

corresponding class. The shape of data points is sigmoid and therefore we fit them with a Gaussian function, which is given in Equation (2). The parameters are described in Li (2011).

$$Y = A_1 + \frac{A_2}{\sqrt{2\Pi}W} \exp - \frac{(X - X_0)^2}{2W^2} \,. \tag{2}$$

3 SLIDING WINDOW TECHNIQUE

In most forecast methods, the prediction of a solar flare is based on the current information about characteristics of the sunspot group or magnetic properties of an active region. The influence of previous parameters is not considered. Some works show that previous information is also very important for short-term solar flare prediction (Yu et al. 2009; Li 2011). It can help improve the performance of a forecasting system. In order to add the information about evolution, the sliding window method is used in our study to convert the original data to a sequence of data.

The sliding-window method can translate values of a predictor at time τ to a data sequence of time, which is represented as

$$x(\tau), x(\tau - \Delta t), ..., x(\tau - w\Delta\tau), I_{\rm tot}(\tau + F) \,. \tag{3}$$

where $x(\tau)$ is the vector of predictors at time τ, and x is the predictors {Area, McIntosh class, Magnetic class, flux}. They are used to forecast whether or not flares will happen within $\tau + F$. F is the forecasting time. $I_{\rm tot}(\tau + F)$ is the total importance of flares within the interval F. The span between $\tau - w\Delta\tau$ and τ is called the sliding window, where w is the window size and $\Delta\tau$ is the interval between two observations. A parameter is extended to w + 1 dimensions by using the sliding window.

Because sunspot data are chosen for each day, the time interval $\Delta\tau$ is 24 h. w is set to 2 for each measure. Generally speaking, the evolution of a sunspot active region lasts for about $3 \sim 5$ d, so it is reasonable to use data observed over 3 d to forecast the flare level. There are not enough data for the sliding window at the beginning of the observation of an active region, so the first observational value is repeated w times to provide the initial values.

4 NEURAL NETWORK METHODS

The machine learning method has been used in solar activity forecasting (Huang et al. 2010, 2012, 2013; Huang & Wang 2013). In our work, two neural network methods were used to build the forecasting model. One is a multilayer perceptron (MLP) network, and the other is a learning vector quantization (LVQ) network. They are described as follows.

4.1 Multilayer Perceptron Network

An MLP is a type of feed-forward network based on the back-propagation learning rule (Witten et al. 2011). The feed-forward network consists of three layers: input layer, hidden layer and output layer. The hidden layer is between the input and the output layers. Weights are connected from an input unit to a hidden unit and from the hidden unit to the output unit. The topology of an MLP network is shown in Figure 1.

Supposing x_i is the input node, w_{ij} is the weight connecting the ith input node with the jth hidden node, w_j is the weight connecting the jth hidden node with the output node and O is the output value, the learning algorithm is divided into two steps. The first step is to calculate the value of nodes in the hidden layer and output layer using the following formulas:

$$h_j = f\left(\sum_{i=1}^{k} w_i^j x_i\right) \,, \tag{4}$$

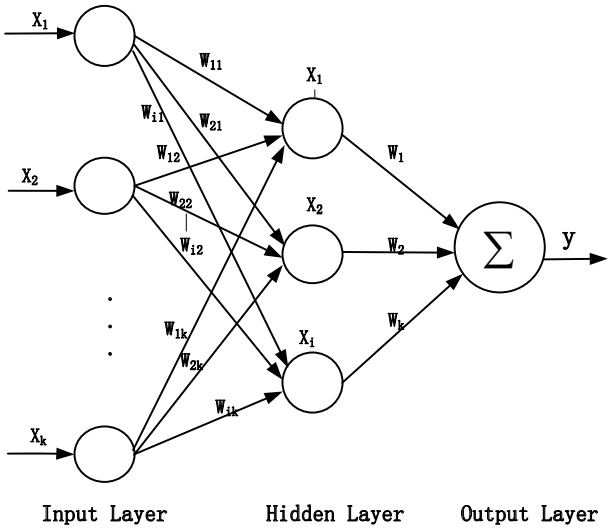

Fig. 1 Structure of an MLP network.

$$y = f\left(\sum_{j=1}^{i} w_i h_j\right) . \tag{5}$$

The sigmoid function is taken as an activation function, which is defined as

$$f(x) = \frac{1}{1+\mathrm{e}^{-x}} . \tag{6}$$

The second step is to conduct the back-propagation training. In this process, the weights are calculated by decreasing the mean square error (MSE). For a given input in the training set, the MSE between the actual output and the desired output needs to be computed. The error function E is represented as

$$E = \frac{1}{2}(T-Y)^2 , \tag{7}$$

where T is the value of the desired target. To reduce the mean square error, it is necessary to calculate the gradient of the error function with respect to each weight. Then each weight is moved in the opposite direction to the gradient. The gradient function for the weights in the output layer is shown below,

$$\delta_j = \frac{\partial E}{\partial w_j} = \frac{\partial (T-Y)^2}{\partial w_j} , \tag{8}$$

$$w_{j+1} = w_j - \eta' \delta_j . \tag{9}$$

From Equation (9), the network weights are updated by multiplying the negative gradient with a step size of the learning rate parameter η. The weights in the hidden layer are calculated with the same procedure. This process repeats the steps using Equations (4) to (9) until the error function E is sufficiently small. In this way, the correct weights are obtained, and the neural network is completely constructed.

In our application, the inputs of the MLP network are the predictors, and the output is the associated flare. If the output value is larger than the threshold 0.5, the output class is a flare and labeled +1; otherwise, the output class is no flare and labeled -1.

4.2 Learning Vector Quantization Network

An LVQ (Kohonen 2001) is one type of neural network, which is based on a competitive learning criterion. The network consists of two layers: an input layer and an output layer. Input units fully connect to output units by weight values. The weight vector associated with each output unit is also called a codebook vector. Similar to other neural network methods, the most important work done by an LVQ is to calculate these weight values. As a competitive network, only the weight vector connected with the winning unit is modified. The winning unit is defined as the closest output unit to the input vector.

Usually several codebook vectors are assigned to each class of x values and x is then assigned to the same class to which the nearest w_i belongs. Define

$$c = \arg\min_i \{\| x - w_i \|\} \tag{10}$$

as a label which has the nearest weight to x, denoted by w_c. Let $x(t)$ be an input sample and let w_i represent the ith weight vector. The LVQ algorithm then checks the input classes against the weight classes and moves w_i appropriately:

(1) If input x and the associated w_c have the same class label, then move them closer together by

$$\Delta w_c(t) = \eta(t)[x(t) - w_c(t)]\,. \tag{11}$$

(2) If input x and the associated w_c have different class labels, then move them apart by

$$\Delta w_c(t) = -\eta(t)[x(t) - w_c(t)]\,. \tag{12}$$

(3) Weights w_i corresponding to other input regions are left unchanged with

$$\Delta w_i(t) = 0, \tag{13}$$

where $\eta(t)$ is the learning rate which satisfies $0 < \eta(t) < 1$. It may be constant or decrease monotonically with time.

The structure of an LVQ network in the proposed model is shown in Figure 2. The input vector of the network is the values of predictors and the output of the network is the possible class label of samples. Once all the weight values are given, the model is constructed.

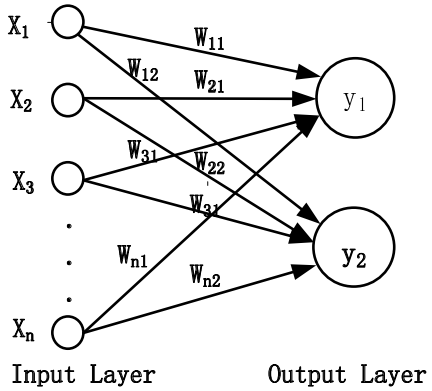

Fig. 2 Structure of an LVQ network.

5 IMPLEMENTATION AND RESULTS

5.1 Implementation

The data set includes 21 655 samples spanning the time from 1996 April 15 to 2008 December 12. The initial sunspot data from the observatory are not suitable to be taken directly as inputs of the forecasting model. There are two steps that need to be done in our work. Firstly, the initial variables are computed on a normalized map by applying the equation describing flare productivity in Section 2, given by Equation (2). Secondly, the normalized data are added to the sequence data by using the sliding window technique described in Section 3. The final sequence of data forms the input of the dataset. The output refers to a classification divided by the importance of solar flares occurring within the coming 48 h, which can yield two cases: greater than or equal to M corresponding to the label of +1, and less than M corresponding to the label of −1.

The input and output data constitute the data set. The data set includes 1252 flaring samples and 20 404 non-flaring samples. It can be found that the number of non-flaring samples is obviously higher than that of flaring samples in the dataset. This is a case of the class imbalance problem that arises in the field of data mining. To solve this problem, an unsupervised clustering method is used in our work. By clustering the sample that does not contain flares with the same number of clusters as the sample with flares, a balanced data set can be obtained. Li (2011) gives a detailed description about applying the clustering algorithm. Finally, the balanced data set is put into the MLP and LVQ algorithms to build the forecasting model. A general schematic view of the system's structure is shown in Figure 3.

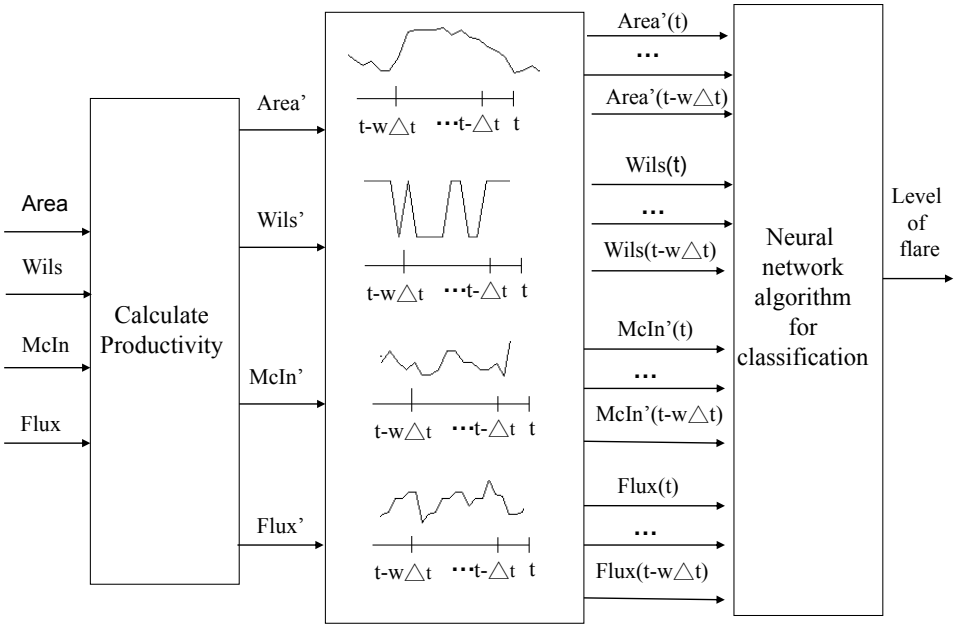

Fig. 3 The workflow of the flare forecasting system. Here Area, Wils, McIn, Flux are the initial values of predictors, and Area', Wils', McIn', Flux' are normalized values using the productivity calculation.

Table 1 Different Outcomes for the Prediction of the Two Classes

	Forecasting Positive	Forecasting Negative
Observation Positive	TP	FN
Observation Negative	FP	TN

5.2 Performance Evaluation

To evaluate the performances of the proposed method, three indexes are used: TP rate, TN rate and Correctness. In solar flare prediction, the samples with flares are labeled as the positive class. Otherwise, they are labeled as the negative class. The TP rate and TN rate are used to evaluate the accuracy of "flaring" and "non-flaring," and Correctness reflects the total accuracy of both. Parameters can be derived from the outcome table for the experiment. In this case, the prediction model has four different possible outcomes, as shown in Table 1. The assumed two classes of samples are denoted as "positive" and "negative," respectively. Samples correctly classified as positive are defined as True Positive (TP); samples correctly classified as negative are defined as True Negative (TN). Samples wrongly predicted as positive are defined as False Positive (FP) and samples wrongly predicted as negative are defined as False Negative (FN).

The TP rate is defined as the ratio of the number of positive class samples predicted as positive to the number of actual positive class samples

$$\mathrm{TP\ rate} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}\,. \tag{14}$$

The TN rate is defined as the ratio of the number of negative class samples predicted as negative to the number of actual negative class samples

$$\mathrm{TN\ rate} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\,. \tag{15}$$

The Correctness is defined as the ratio of the number of positive class samples predicted as positive to the number of actual positive class samples, and for negative samples as well

$$\mathrm{Correctness} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}\,. \tag{16}$$

5.3 Experimental Results and Analyses

The data set is divided into 10 folds. Therein nine folds are used for training and the remaining one fold for testing. The training set is applied in the algorithms described in Section 4 to build the forecasting model. When the testing set is input into the models, the forecasting results are obtained. This process is repeated 10 times, and the average value of test accuracy is considered as an estimation of the prediction performance.

The proposed flare forecasting algorithms are implemented in the Waikato Environment for Knowledge Analysis (WEKA), which is data mining software written in Java (Witten et al. 2011). WEKA can be freely downloaded from *http://www.cs.waikato.ac.nz/ml/weka/*. In the MLP algorithm, the number of hidden layer nodes and the learning are kept at the default settings. In the LVQ algorithm, the number of codebook vectors is set to 150 and the number of total training iterations is set to 50.

To validate the availability of the sliding window technique, window size w is set to two values: 0 and 2 for the two algorithms. The value of 0 means that the dataset keeps its original type and no sequential data are added; 1 means that three days of sequence data are introduced into the

Table 2 Comparison of Sliding window Size Using the MLP Method

Experiment	$w = 0$			$w = 2$		
	TP rate (%)	TN rate (%)	CORR (%)	TP rate (%)	TN rate (%)	CORR (%)
1	76	82.5	78.93	66.4	88.4	87.1
2	77.6	77.9	77.92	72.8	83	82.4
3	78.4	79	78.94	66.4	86.1	84.92
4	53.6	94	91.6	56	91.7	92.75
5	41.6	95	91.92	51.2	93.4	90.95
6	68	82.5	81.71	68	79.5	78.85
7	76	78.8	78.67	70.4	79.1	78.56
8	58.4	86.7	85.03	60.8	86.8	85.27
9	89.6	77	77.7	86.4	81.1	81.43
10	78.4	78.2	78.25	80.8	80.2	80.28
Average	69.76	83.16	82.39	67.92	84.93	84.25

Table 3 Comparison of Sliding window Size Using the LVQ Method

Experiment	$w = 0$			$w = 2$		
	TP rate (%)	TN rate (%)	CORR (%)	TP rate (%)	TN rate (%)	CORR (%)
1	74.4	81.97	81.48	71.2	83.33	82.63
2	72	77.75	77.41	72	80.15	79.68
3	66.4	86.67	85.5	64.8	87.4	86.1
4	52.8	93.48	91.13	56.8	93.87	91.73
5	62.4	84.8	83.51	56	90.59	88.59
6	76.8	72.71	72.47	63.2	81.3	80
7	79.2	71.32	71.78	65.6	80.05	79.21
8	69.6	77.06	76.63	62.4	84.07	82.82
9	88	70.25	71.27	86.4	80.15	81.51
10	86.4	65.34	66.56	80	76.52	76.72
Average	72.8	78.13	77.77	67.84	83.74	82.80

dataset. Ten experiments are carried out and results for MLP and LVQ are shown in Tables 2 and 3 respectively.

From Tables 2 and 3, we can see that the total prediction accuracy (CORR) has improved with predictors related to evolution in contrast to the original ones (MLP: from 82.39% to 84.25%, LVQ: from 77.77% to 82.80%). This is mainly reflected in the TN rate. The TN rate for MLP is 83.16% with a window size of 0 and 84.93% with a window size of 2. The performance of the LVQ algorithm has improved more obviously. The TN rate of LVQ has changed from 78.13% to 83.74% (an increase of 5.61%) with increasing widow size. This means that the information about evolution of the predictors introduced by the sliding window is effective, especially for forecasting the case of no flare occurring, and the MLP and LVQ can extract this information.

6 CONCLUSIONS

Studies have found that not only does flare occurrence depend on characteristics of the current morphological and photospheric magnetic field, but it is also influenced by previous properties. To verify the effectiveness of information about evolution in the sunspot data of an active region, a short-term solar flare prediction model is established using sequential sunspot data. The prediction model is established on data spanning the time from January 1996 to December 2008. The information about sunspot evolution covering three days or 12 dimensions of the input vector is introduced into neural network algorithms by using a sliding window method, and MLP and LVQ are employed to learn prediction models from this information. The experimental results show that short-term prediction of

solar flares within the sequential data describing evolution of sunspots is effective. Compared with the model with sequential data describing the magnetic field proposed by Li (2011), the prediction accuracy is 8.2% (82.80%: 74.58%) higher than in the LVQ method based on data describing the magnetic field.

Some improvements can be made in future work. The experimental results show that predictors describing the evolution of sunspots are not enough to forecast "flare" events. In order to improve the accuracy of forecasting both "flare" and "non flare" events, the predictors for sunspots can be connected with those describing magnetic fields to build an integrated flare forecasting model in the future.

Acknowledgements We thank the reviewer for his valuable comments which helped to considerably improve the quality of the manuscript. This work is supported by the National Natural Science Foundation of China (Grant Nos. 10973020 and 11273031).

References

Gallagher, P. T., Moon, Y.-J., & Wang, H. 2002, Sol. Phys., 209, 171
Han, Z. Z., Zhou, S. R., & Wu, Q. D. 2001, Science in China G: Physics and Astronomy, 31, 274
Huang, X., Yu, D., Hu, Q., Wang, H., & Cui, Y. 2010, Sol. Phys., 263, 175
Huang, X., Wang, H., & Dai, X. 2012, Science in China G: Physics and Astronomy, 55, 1956
Huang, X., Zhang, L., Wang, H., & Li, L. 2013, A&A, 549, A127
Huang, X., & Wang, H.-N. 2013, RAA (Research in Astronomy and Astrophysics), 13, 351
Kohonen, T. 2001, Self-Organizing Maps (New York: Springer)
Li, R. 2011, Scientia Sinica Physica, Mechanica & Astronomica, 41, 1342
Li, R., Wang, H.-N., He, H., Cui, Y.-M., & Zhan-LeDu 2007, ChJAA (Chin. J. Astron. Astrophys.), 7, 441
McIntosh, P. S. 1990, Sol. Phys., 125, 251
Qahwaji, R., & Colak, T. 2007, Sol. Phys., 241, 195
Wang, H. N., Cui, Y. M., Li, R., Zhang, L. Y., & Han, H. 2008, Advances in Space Research, 42, 1464
Wang, H.-N., Cui, Y.-M., & He, H. 2009, RAA (Research in Astronomy and Astrophysics), 9, 687
Wang, H., Ai, G., & Wang, J. 2012, Space Weather, 10, S09003
Wang, J. L., & Zhang, G. Q. 1994, Progress in Geophysics (in Chinese), 9 (Suppl.), 1
Wheatland, M. S. 2005, Space Weather, 3, 07003
Witten, I. H., Frank, E., & Hall, M. A. 2011, Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques (Morgan Kaufmann)
Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, Sol. Phys., 255, 91
Zhu, C. L., & Wang, J. L. 2003, ChJAA (Chin. J. Astron. Astrophys.), 3, 563