Hindawi
Advances in Astronomy
Volume 2019, Article ID 5190353, 6 pages
https://doi.org/10.1155/2019/5190353

Research Article

Full-Disk Solar Flare Forecasting Model Based on Data Mining Method

Rong Li[1] **and Yong Du**[2]

[1]*School of Information, Beijing Wuzi University, Beijing 101149, China*
[2]*Department of Electrical and Information Engineering, Northeast Agricultural University, Harbin, China*

Correspondence should be addressed to Rong Li; lirong@bao.ac.cn

Received 11 April 2019; Accepted 18 June 2019; Published 1 August 2019

Guest Editor: Liyun Zhang

Solar flare is one of the violent solar eruptive phenomena; many solar flare forecasting models are built based on the properties of active regions. However, most of these models only focus on active regions within 30° of solar disk center because of the projection effect. Using cost sensitive decision tree algorithm, we build two solar flare forecasting models from the active regions within 30° of solar disk center and outside 30° of solar disk center, respectively. The performances of these two models are compared and analyzed. Merging these two models into a single one, we obtain a full-disk solar flare forecasting model.

1. Introduction

Solar activities are the primary source of space weather. As one of the important solar eruptive phenomena, solar flares associated with the electromagnetic radiation and energetic particles often interfere with geostationary satellites, communication systems, and even power grids [1]. Therefore, solar flare forecasting is a significant topic in space weather forecasting community.

Because the trigger mechanisms of solar flares are unsolved, the current solar flare forecasting only depends on the probabilistic model. The statistical and data mining methods are used to build solar flare forecasting models. Miller (1989) developed an expert system (WOLF) to forecast the probable occurrence of solar flares [2]. McIntosh (1990) summarized the McIntosh classifications of sunspots and built an expert system (Theo) to forecast X-ray flares [3]. Long after this work, the McIntosh classifications are considered as a guide in forecasting solar flares in many space weather prediction centers. Measuring contributions of the McIntosh classifications for solar flare forecasting, Bornmann and Shaw (1994) built a solar flare forecasting model using multiple liner regression analysis [4]. Wheatland (2004) pointed out that the history of solar flares is also an important indicator for the occurrence of solar flares, so a Bayesian approach was proposed to forecast solar flares using the previous flaring records [5]. Leka and Barnes (2007) applied discriminant analysis to produce a binary categorization for the flaring and nonflaring regions [6], and this approach was extended to a probabilistic forecast in Barnes et al. (2007) [7].

Data mining methods also have a long history for the application in solar flare forecasting. Bradshaw et al. (1989) trained a three-layer neural network to forecast flares [8]. Wang et al. (2008) built a solar flare forecasting model supported with an artificial neural network based on the solar magnetic field parameters [9]. Li et al. (2007) proposed a data mining method combining the support vector machine and the k-nearest neighbors to train a solar flare forecasting model [10]. Qahwaji and Colak (2007) built a hybrid system that combines a support vector machine and a cascade-correlation neural network for solar flare forecasting [11]. The sequential information of active regions is analyzed in [12–16]. The active longitudes information is used to improve the performance of solar flare forecasting in [17]. At present, deep learning methods have been used to build solar flare forecasting models [18, 19].

Because of the projection effect of solar magnetograms, active regions within 30° of solar disk center, where projection effect can be negligible, are usually selected to extract parameters and furthermore to build the forecasting model.

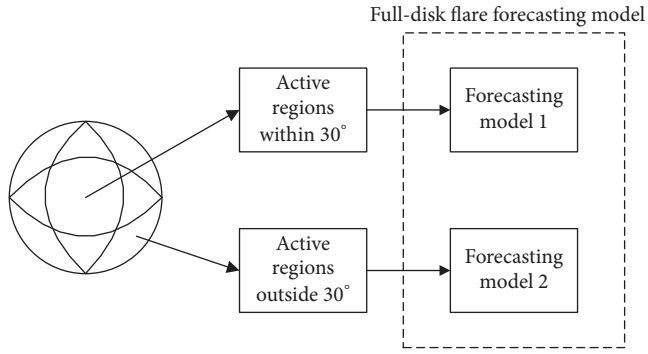

Figure 1: Full-disk solar flare forecasting model.

However, active regions which locate outside 30° of solar disk center also produce solar flares. In the present work, we collect the data for active regions outside 30° of solar disk center and their related solar flares and build a solar flare forecasting model from this dataset. Combining it with the solar flare forecasting model trained from active regions within 30° of solar disk center, we obtain a full-disk solar flare forecasting model shown in Figure 1.

The paper is organized as follows. In Section 2, we introduce active region parameters and the related flare catalog. In Section 3, we describe the data mining method. In Section 4, we estimate the performance of the solar flare forecasting model. And finally, in Section 5, we give a brief summary of this work.

2. Data

2.1. Active Region Data. The Solar Dynamics Observatory (SDO) satellite is launched on 2010 February. The Helioseismic and Magnetic Imager (HMI), which is one of three instruments aboard the SDO, measures the full-disk photospheric vector magnetic field [20]. In 2014, a data product called Space Weather HMI Active Region Patches (SHARP) automatically identifies active regions using the vector magnetic field data when these active regions cross the solar disk [21]. For this study, we use the active region vector magnetic field data generated by the SDO's SHARP data patches from 2011 August to 2012 July. We calculate 4 physical parameters using these 12 month vector magnetic field data, and obtain 2966 samples including 1436 samples within 30° of solar disk center and 1530 samples outside 30° of solar disk center.

The 4 physical parameters are:

(1) The maximum horizontal gradient of the longitudinal magnetic field: this parameter estimates maximum squeezing among flux systems in an active region.

(2) The length of neutral lines: the neutral lines separate opposite polarities of the longitudinal magnetic field [22].

(3) The number of singular points: it is the number of nodes in the network formed by magnetic separatrices [22].

(4) Sum of photospheric magnetic free energy.

$$\rho_{\text{sum}} = \sum \left(B^{obs} - B^{pot}\right)^2 \quad (1)$$

ρ_{sum} measures the nonpotentiality of an active region.

2.2. Flare Data. According to the peak flux of 1 to 8 angstrom X-rays, solar flare is classified as different class levels shown in Table 1. Within a class level, there is a linear scale from 1 to 9. For example, a C2 flare is twice as powerful as a C1 flare.

Solar flares whose Geostationary Operational Environmental Satellite (GOES) X-ray flux peak magnitude is larger than the C1.0 level are considered in the present work. Solar flare data is collected from the National Geophysical Data Center GOES X-ray flux flare catalogs. An active region is considered as a flaring sample, when this region produces a flare whose level is larger than C1.0 within 48 hours after the observation of this active region. Otherwise, an active region is considered as a nonflaring sample. As such, there are 74 flaring samples and 1362 nonflaring samples for active regions within 30° of solar disk center. And there are 101 flaring samples and 1429 nonflaring samples for active regions outside 30° of solar disk center.

3. Method

3.1. Basic Algorithm. A decision tree is a flowchart-like model that shows the various outcomes from a series of decisions. It can be used for research analysis or for building forecasting model.

Decision trees have three main parts: a root node, leaf nodes, and branches. The root node is the starting point, root contains questions or criteria to be answered, and leaf nodes stand for the decision of the model. Branches are arrows connecting nodes, showing the information flow between the nodes.

The decision tree algorithm is used to build the solar flare forecasting model. This means that the forecasting model will be represented by a tree-like structure shown in Figure 2 [23]. The decision tree consists of testing nodes, leaf nodes, and branches. A sample is classified from the root node. The specified parameter of this node is calculated and the sample is moved down along the corresponding branch and

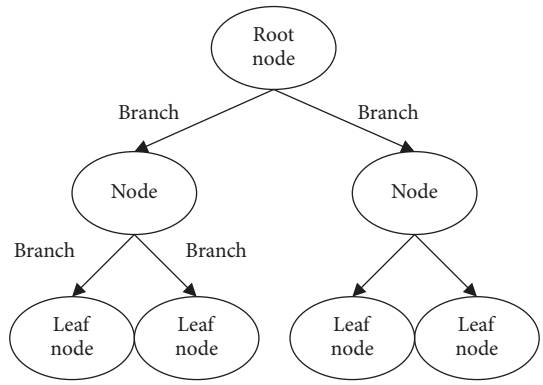

Figure 2: Structure of decision tree.

Table 1: Classifications of solar X-ray flares.

Class level	peak flux of 1 to 8 angstrom X-rays (Watts/square metre)
A	$< 10^{-7}$
B	10^{-7} - 10^{-6}
C	10^{-6} - 10^{-5}
M	10^{-5} - 10^{-4}
X	$> 10^{-4}$

finally goes to the leaf node where the classification result is determined.

The decision tree is constructed from the training set recursively. In each step, the best parameter is selected to generate the test node and the corresponding branches. The parameter is evaluated by information gain rat

$$GR(D,F) = \frac{IG(D,F)}{H(F)} \tag{2}$$

where D stands for the decision of the model, F stands for the feature of the model, IG(D,F)=H(D)-H(D|F) is the information gain (IG), and H stands for the entropy which is used to measure the uncertainty of a system.

The training dataset is divided into some subsets according to the value of branches. This process is repeated until the following stop criteria are satisfied: (1) samples in the subset have the same class label or (2) all possible tests have the same class distribution [24]. When the stop criteria are satisfied, the leaf node is generated. The class label of the samples in the leaf node is the same as that of the majority of samples in this leaf node.

3.2. Cost Sensitive Modification for the Basic Algorithm. As shown in Section 2, the ratio between nonflaring samples and the flaring samples is 16. This is called class imbalance problem in data mining community. In order to treat the class imbalance problem, we modified the basic algorithm to the cost sensitive one [25].

In the basic decision tree algorithm, the probability is a basic component to calculate the entropy, information gain, and information gain ratio. Generally, the probability is estimated by the frequency calculated from the dataset.

$$P(D = d_i) = \frac{|D = d_i|}{|D|} \tag{3}$$

where |D| is the number of samples in set D and $|D = d_i|$ is the number of samples with class label d_i in set D.

In the cost sensitive algorithm, there are different costs for different class labels. For example, for a binary classification problem, the cost for class d_0 is C_0, and the cost for class d_1 is C_1. Thus, the probability for cost sensitive problem can be estimated as follows.

$$P_{cost}(D = d_0) = \frac{|D = d_0| \times C_0}{|D = d_0| \times C_0 + |D = d_1| \times C_1} \tag{4}$$

$$P_{cost}(D = d1) = \frac{|D = d_1| \times C_1}{|D = d_0| \times C_0 + |D = d_1| \times C_1} \tag{5}$$

In fact, the usual probability is considered as the cost sensitive probability when the costs C_0 and C_1 are settled to 1. Using the cost sensitive probability, we can calculate the cost sensitive entropy and information gain. Similar to the basic decision tree algorithm, we can build the cost sensitive decision tree model.

4. Performance

4.1. Performance Metrics. For a binary forecasting model, the results can be summarized in contingency table shown in Table 2. The flaring sample is called positive one, and the nonflaring sample is called negative one. The actual positive sample correctly forecasted as positive one is called true

Table 2: Definition of contingency table.

	Forecast positive	Forecast negative
Actual positive	N_{TP}	N_{FN}
Actual negative	N_{FP}	N_{TN}

Table 3: Contingency table for model 1.

	Forecast positive	Forecast negative
Actual positive	53	21
Actual negative	161	1201

Table 4: Contingency table for model 2.

	Forecast positive	Forecast negative
Actual positive	78	23
Actual negative	464	965

Table 5: Performances of solar flare forecasting models.

	TP rate	TN rate	HSS
Model 1	71.6%	88.2%	0.316
Model 2	77.2%	67.5%	0.148

positive (TP), the actual positive sample wrongly forecasted as negative one is false negative (FN), the actual negative sample correctly forecasted as negative one is true negative (TN), and the true negative sample wrongly forecasted as positive one is false positive (FP).

Using the contingency table, 3 performance metrics are defined to compare the performance of the forecasting model. The TP rate and TN rate are defined to evaluate the accuracy of flaring samples and nonflaring samples, respectively.

$$TPrate = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{6}$$

$$TNrate = \frac{N_{TN}}{N_{TN} + N_{FP}} \tag{7}$$

Heidke skill score (HSS) is used to evaluate the increase in forecasting power over that of random forecast:

$$HSS = \frac{PC - E}{1 - E} \tag{8}$$

where $PC = (N_{TP} + N_{TN})/(N_{TP} + N_{TN} + N_{FN} + N_{FP})$ and

$$E = \frac{(N_{TP} + N_{FN})(N_{TP} + N_{FP})}{(N_{TP} + N_{TN} + N_{TP} + N_{FP})^2} + \frac{(N_{TN} + N_{FP})(N_{TN} + N_{FN})}{(N_{TP} + N_{TN} + N_{FN} + N_{FP})^2}. \tag{9}$$

4.2. Results. There are 2966 samples in the dataset. In order to make good use of this data, leave-one-out cross validation method is used to evaluate the performance of the forecasting model. In this method, all but one of the samples is used as training set, and only one sample is used as testing set. The process is repeated as many times as the number of samples in the dataset. Leave-one-out cross validation method does not waste data; however, it is computationally expensive.

Cost sensitive decision tree is an efficient algorithm, so we can complete the leave-one-out testing. The cost for flaring samples is 50 times larger than that for nonflaring samples.

In order to simplify the following discussion, solar flare forecasting model learned from samples within the 30° of solar disk center is called model 1. And solar flare forecasting model learned from samples outside the 30° of solar disk center is called model 2. The contingency tables of model 1 and model 2 are shown in Tables 3 and 4. Based on these contingency tables, the performances of the two forecasting models can be compared by the performance metrics shown in Table 5.

From Table 5, we can find that the performance of model 2 is worse than that of model 1, because the physical parameters used in model 2 could be influenced by the projection effect. However, the performance of model 2 is acceptable. Combining model 1 and model 2, we can obtain a full-disk solar flare forecasting model.

At present, little work can provide forecasting results of solar flares in the active region beyond 30 degrees of the solar disk; hence, we choose the forecasting results in the active region within 30 degrees to compare them with the flare forecasting results provided by the convolution neural network [24]. The results are shown in Table 6. We find that the flare forecasting model built by the convolution neural network has a higher TP rate, while our forecasting model has a higher TN rate. Because the proportions of flaring samples and nonflaring samples are different in the testing dataset, the HSS is incomparable.

Table 6: Performance comparisons.

Performance index	Decision tree	CNN
TP rate	71.6%	85.0%
TN rate	88.2%	81.0%
HSS	0.316	0.143

5. Conclusion

Space Weather HMI Active Region Patches data product automatically identifies the active regions when they cross the solar disk. We classify the active region samples into two groups by their location information. The active region samples located within the 30° of solar disk center are classified into group one, and the rest of samples are classified into group two. The projection effect of the samples in group one can be negligible, but the magnetic parameters extracted from active region in group two could not be too accurate because of the projection effect. Two solar flare forecasting models are built using data mining method from two group samples, respectively. The performances of these two forecasting models are estimated. The performance of the forecasting model learned from samples within the 30° of solar disk center is better than that of the forecasting model learned from other samples, because the parameters extracted from the active regions outside the 30° of solar disk center are not accurate enough, and the uncertainty is introduced to evaluate the nonpotentiality of these active regions. A full-disk solar flare forecasting model is generated by combining the two models together.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The data used herein was made possible by funding to NWRA from NASA/LWS contract NNH09CE72C (Dr. Graham Barnes, PI). This work is supported by the National Natural Science Foundation of China (NSFC) (Grant No. 11303051), Beijing Intelligent Logistics System Collaborative Innovation Center, and Beijing Key Laboratory (No. BZ0211).

References

[1] G. Ai, H. Wang, and J. Wang, "What is a solar electromagnetic storm?" *Space Weather Journal*, vol. 10, no. 9, 2012.

[2] A. Heck and F. Murtagh, "Knowledge-based systems in astronomy," in *Lecture Notes in Physics*, 1989, 329.

[3] P. S. McIntosh, "The classification of sunspot groups," *Solar Physics*, vol. 125, no. 2, pp. 251–267, 1990.

[4] P. L. Bornmann and D. Shaw, "Flare rates and the McIntosh active-region classifications," *Solar Physics*, vol. 150, no. 1-2, pp. 127–146, 1994.

[5] M. S. Wheatland, "A bayesian approach to solar flare prediction iop-2016.pngA publishing partnershipA Bayesian Approach to Solar Flare Prediction," *The Astrophysical Journal*, vol. 609, p. 1134, 2004.

[6] K. D. Leka and G. Barnes, "Photospheric magnetic field properties of flaring versus flare-quiet active regions. iv. a statistically significant sample," *The Astrophysical Journal*, vol. 656, no. 2, p. 1173, 2007.

[7] G. Barnes, K. D. Leka, E. A. Schumer, and D. J. Della-Rose, "Probabilistic forecasting of solar flares from vector magnetogram data," *Space Weather*, vol. 5, no. 9, p. S09002, 2007.

[8] G. Bradshaw, R. Fozzard, and L. Ceci, "A connectionist expert system that actually works," *Adv Neu Inform Proc Sys*, vol. 1, pp. 248–255, 1989.

[9] H. N. Wang, Y. M. Cui, R. Li, L. Y. Zhang, and H. Han, "Solar flare forecasting model supported with artificial neural network techniques," *Advances in Space Research*, vol. 42, no. 9, p. 1464, 2008.

[10] R. Li, H. N. Wang, H. He, Y. M. Cui, and Z. L. Du, "Support vector machine combined with k-nearest neighbors for solar flare forecasting," *Chinese Journal of Astronomy and Astrophysics*, vol. 7, no. 3, p. 441, 2007.

[11] R. Qahwaji and T. Colak, "Automatic short-term solar flare prediction using machine learning and sunspot associations," *Solar Physics*, vol. 241, p. 195, 2007.

[12] D. Yu, X. Huang, H. Wang, and Y. Cui, "Short-term solar flare prediction using a sequential supervised learning method," *Solar Physics*, vol. 255, no. 1, pp. 91–105, 2009.

[13] D. Yu, X. Huang, Q. Hu, R. Zhou, H. Wang, and Y. Cui, "Short-term solar flare prediction using multiresolution predictors," *The Astrophysical Journal*, vol. 709, no. 1, p. 321, 2010.

[14] D. Yu, X. Huang, H. Wang, Y. Cui, Q. Hu, and R. Zhou, "Short-term solar flare level prediction using a bayesian network approach," *The Astrophysical Journal*, vol. 710, no. 1, p. 869, 2010.

[15] X. Huang, D. Yu, Q. Hu, H. Wang, and Y. Cui, "Short-term solar flare prediction using predictor teams," *Solar Physics*, vol. 263, no. 1-2, pp. 175–184, 2010.

[16] X. Huang and H. N. Wnag, "Solar flare prediction using highly stressed longitudinal magnetic field parameters," *Research in astronomy and astrophysics*, vol. 13, no. 3, pp. 351–358, 2013.

[17] X. Huang, L. Zhang, H. Wang, and L. Li, "Improving the performance of solar flare prediction using active longitudes information," *Astronomy & Astrophysics*, vol. 549, article A127, p. 6, 2013.

[18] X. Huang, H. Wang, L. Xu, J. Liu, R. Li, and X. Dai, "Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms," *The Astrophysical Journal*, vol. 856, no. 1, p. 7, 2018.

[19] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, and M. Ishii, "Deep flare net (defn) model for solar flare prediction," *The Astrophysical Journal*, vol. 858, no. 2, 2018.

[20] J. Schou, P. H. Scherrer, R. I. Bush et al., "Design and ground calibration of the helioseismic and magnetic imager (hmi) instrument on the solar dynamics observatory (sdo)," *Solar Physics*, vol. 275, no. 1-2, pp. 229–259, 2012.

[21] M. G. Bobra, X. Sun, J. T. Hoeksema et al., "The helioseismic and magnetic imager (hmi) vector magnetic field pipeline: sharps – space-weather hmi active region patches," *Solar Physics*, vol. 289, no. 9, pp. 3549–3578, 2014.

[22] Y. Cui, R. Li, L. Zhang, Y. He, and H. Wang, "Correlation between solar flare productivity and photospheric magnetic field properties," *Solar Physics*, vol. 237, p. 45, 2006.

[23] X. Huang, H. N. Wang, and X. H. Dai, "Science china physics," *Mechanics and Astronomy*, vol. 55, no. 10, pp. 1956–1962.

[24] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, Calif, USA, 1993.

[25] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, San Mateo, Calif, USA, 2005.