Research in Astron. Astrophys. **2012** Vol. **12** No. **9**, 1191–1196
http://www.raa-journal.org *http://www.iop.org/journals/raa*

Research in Astronomy and Astrophysics

LETTERS

Automated flare prediction using the AdaBoost algorithm *

Ru-Shi Lan[1], Yong Jiang[1], Liu-Guan Ding[2] and Jian-Wei Yang[1]

[1] School of Math and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China; *lrs0106@yahoo.com, jiang@nuist.edu.cn*
[2] School of Physics and Optoelectronic Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China

Received 2012 February 23; accepted 2012 July 2

Abstract We propose a flare prediction method based on the AdaBoost algorithm, which constructs a strong prediction model from a combination of several basic models. Three predictors, extracted from the photospheric magnetograms, are applied as features to predict the occurrence of flares with a certain level over 24 hours following the time when the magnetogram is recorded. To demonstrate the effectiveness of the proposed method, comparisons of experimental results with respect to some existing methods are given. The results show that an improvement is achieved in predicting the occurrences of large flares.

Key words: Sun: flares — Sun: magnetic fields

1 INTRODUCTION

Solar flares are the result of a sudden and intense release of magnetic energy stored in the solar corona (Dauphin et al. 2007; Du & Wang 2012). The released energy, in the form of electromagnetic emissions and particle acceleration, can have an important influence on communication systems, space based systems, and even human life or health. Therefore, it is a significant task to understand the mechanism of solar flares and forecast them. Several mechanisms have been developed to explain the occurrence of solar flares, such as flux emergence and cancellation (Gan et al. 1993; Zhang et al. 2001), kink instability of coronal flux ropes (Sakurai 1976; Li & Gan 2011), and magnetic reconnection (Forbes et al. 2006; Fang et al. 2010; Fang 2011).

There are two main aspects in flare prediction, namely to construct informative predictors, and to build powerful prediction models. Many predictors have been proposed, and the predictors derived from the magnetic observation have recently aroused a great deal of interest. Of the proposed measurements, the predictors proposed by Jing et al. (2006) and Cui et al. (2006) have been extensively investigated (Song et al. 2009; Yuan et al. 2010; Yu et al. 2009, 2010a,b). On the other hand, flare prediction actually can be regarded as a classification task. As a result, several machine learning approaches have been applied to develop solar flare prediction models, such as Bayesian network (Yu et al. 2010b), neural network (Qahwaji & Colak 2007), k-nearest neighbors (Li et al. 2007), and C4.5 decision tree (Yu et al. 2009). Promising results have been achieved.

∗ Supported by the National Natural Science Foundation of China.

Based on Song et al.'s work (Song et al. 2009), Yuan et al. (2010) proposed a method by combing an ordinal logistic regression (LR) model and a support vector machine (SVM) classifier to predict the occurrence of a certain level of flares within 24 hours. The experimental results inspire us to build a prediction model by combining some existing ones. Therefore, a method, based on the AdaBoost algorithm (Freund & Schapire 1997) is proposed for flare prediction in this paper because Adaboost is a meta-algorithm which can be used in conjunction with many other prediction methods. Comparison of results shows an improvement in the accuracy of M- and X-class flare prediction.

The remainder of the paper is organized as follows. In Section 2, the three predictors used in this paper are briefly reviewed. The proposed method is given in Section 3, and experiments and results are shown in Section 4. In the last section, some conclusions are given.

2 DESCRIPTION OF PREDICTORS

To make a reasonable and fair comparison, the predictors, proposed by Song et al. (2009), are used in this paper. The predictors are composed of:

– Total unsigned magnetic flux, T_{flux},

$$T_{\text{flux}} = \iint |B_z| dx dy, \tag{1}$$

where B_z is the pixel intensity of MDI magnetographs. T_{flux} denotes the integration of pixel intensity over the area of an active region.

– Length of the strong-gradient magnetic polarity inversion line, L_{gpi}, which is measured by the total number of pixels on which the gradient $|\nabla_\perp B_z|$ is greater than $50\ \text{G}\ \text{Mm}^{-1}$. The definition of $|\nabla_\perp B_z|$ is given as

$$|\nabla_\perp B_z| = \left[\left(\frac{dB_z}{dx} \right)^2 + \left(\frac{dB_z}{dy} \right)^2 \right]^{1/2}. \tag{2}$$

– Total magnetic energy dissipation, E_{diss},

$$E_{\text{diss}} = \iint 4 \left[\left(\frac{dB_z}{dx} \right)^2 + \left(\frac{dB_z}{dy} \right)^2 \right] + 2 \left(\frac{dB_z}{dx} + \frac{dB_z}{dy} \right)^2 dx dy, \tag{3}$$

where the integration is performed over the area of an active region.

For more details on the predictors, please refer to Jing et al. (2006) and Song et al. (2009).

3 ADABOOST ALGORITHM

As mentioned before, the proposed prediction method is based on AdaBoost. In this section, we briefly review the AdaBoost algorithm. AdaBoost, derived from Boost, is one of the most important schemes in ensemble learning. It was first introduced by Freund & Schapire (1997), whose main idea was to combine several basic or weak prediction models together to develop a more effective and practical model. In this work, a very simple version of Adaboost is used.

The AdaBoost algorithm takes the following input data set: $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$, where $\boldsymbol{x}_i$ is a predictor vector in the input data belonging to the class y_i, and $y_i \in \{+1, -1\}$. In this paper, the predictor vector consists of T_{flux}, L_{gpi}, and E_{diss} mentioned in Section 2. $y_i = +1$ denotes that there will be an eruption of a flare with the predictor vector $\boldsymbol{x}_i$, and vice versa. In fact, the flare prediction discussed here is equivalent to a binary classification problem. Suppose there are T basic binary classifiers $Z = \{z_1(\boldsymbol{x}), z_2(\boldsymbol{x}), \cdots, z_T(\boldsymbol{x})\}$ which are included in

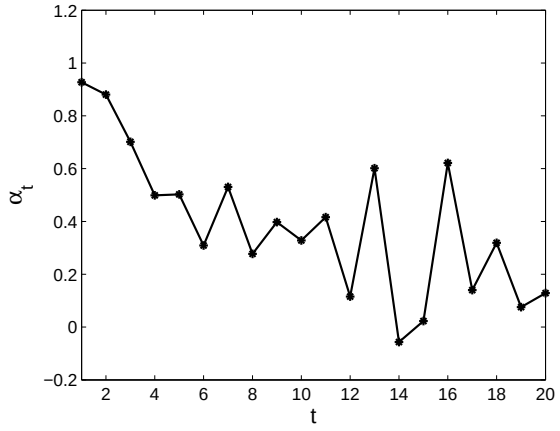

Fig. 1 The weight α_t for each basic classifier.

Adaboost. AdaBoost repeatedly performs one basic classification function over a series of time steps $t = 1, 2, \cdots, T$ in order to calculate a weight α_t for $z_t(\boldsymbol{x})$. α_t represents the contribution of $z_t(\boldsymbol{x})$ to the derived classifier, and is determined by the total prediction errors ε_t of $z_t(\boldsymbol{x})$. Therefore, $z_t(\boldsymbol{x})$ will be given a larger α_t on the condition that ε_t is smaller to improve the accuracy of the derived classifier. As an example, Figure 1 illustrates the weight α_t used in the following experiment. We can observe that α_t is different for each basic classifier. In addition, the input training data are given with equally initialized weights $\omega_1, \omega_2, \cdots, \omega_m$, and these weights will be changed during the training process. The outline of AdaBoost is given as follows.

Input: Training data D, and the basic classifier set Z.

Output: A strong classifier H.

Step 1: Initialize the weights for D, namely $\omega_i = 1/m$;

Step 2: Repeatedly execute the following sub-steps from $t = 1, 2, \cdots, T$,

(1) Calculate the prediction results $\bar{y_i}$ and total error ε_t of z_t:

$$\bar{y_i} = \text{sign}(z_t(\boldsymbol{x_i})), i = 1, 2, \cdots, m, \text{ and } \varepsilon_t = \sum_{i=1}^{m} \omega_i (\bar{y_i} \neq y_i),$$

where sign(.) is the signum function for a nonzero element.

(2) Calculate the weight α_t for z_t:

$$\alpha_t = \frac{1}{2} \log \frac{1 - \varepsilon_t}{\varepsilon_t};$$

(3) Update the weights $\omega_1, \omega_2, \cdots, \omega_m$ for D:

$$\omega_i = \omega_i \exp(-\alpha_t z_t(\boldsymbol{x_i}) \bar{y_i}), \ i = 1, 2, \cdots, m;$$

(4) Normalize ω_i in the following way:

$$\omega_i = \omega_i / \sum_{i=1}^{m} \omega_i;$$

Step 3: The final strong classifier is:

$$H = \text{sign}\left(\sum_{t=1}^{T} \alpha_t z_t(\boldsymbol{x})\right) .$$

For more details on the AdaBoost algorithm, see Freund & Schapire (1997) and Schapire (2003).

4 EXPERIMENTAL RESULTS

In this section, the effectiveness of the proposed method is evaluated by some experiments. The data set (Song et al. 2009) is applied in the paper, which contains 230 active regions extracted from *SOHO*/MDI magnetograms. The flares in these active regions are denoted as level zero to level three according their intensity, such that level zero means that it is flaring quiet or A and/or B class flare, while level three represents the occurrence of at least one X-class flare in the active region. All the three predictors are extracted from the active regions and then used to make a prediction.

We compare the proposed method with LR, SVM, and LR+SVM prediction models. All the methods are tested similarly to Yuan's work. The leave-one-out cross-validation approach is applied in the following experiments, namely for 230 samples, 229 samples are used as training input data, and the remaining one is for testing. As a result, the process is repeated 230 times to test every sample. The prediction results are recorded in Table 1.

Table 1 A Sample Contingency Table

	Observation Positive	Observation Negative
Prediction Positive	a	b
Prediction Negative	c	d

Based on the results given in Table 1, seven measurements are calculated as follows to assess the performance of all the methods from different aspects (Yuan et al. 2010).

- Correctness $= (a+d)/(a+b+c+d)$;
- True Positive $= a/(a+b)$;
- True Negative $= d/(c+d)$;
- Weighted True Rate $= a/(a+b)*(a+c)/(a+b+c+d)+d/(c+d)*(b+d)/(a+b+c+d)$;
- Positive Accuracy $= a/(a+c)$;
- Negative Accuracy $= d/(b+d)$;
- Weighted Accuracy $= a/(a+c)*(a+c)/(a+b+c+d)+d/(b+d)*(b+d)/(a+b+c+d)$.

Tables 2 and 3 show the binary prediction results of all the test methods for level zero to three respectively. Observing the results, we can find that none of the test methods can obtain the best results for all types of flare levels. For level zero, as shown in Table 3, performances of SVM, LR+SVM, and the proposed method are almost the same. In this situation, the Positive Accuracy of the LR-based method is higher than other methods' results, but its Negative Accuracy is lower. When considering level one, as illustrated in Table 3, the LR-based method surpasses the other methods in all aspects.

However, predicting the occurrence of large flares is a more significant and meaningful goal. For flares at level two or three, LR-based methods do not work very well. At the same time, the proposed method outperforms the other three test methods at these levels. Considering level two, the number of correct predictions of flare eruption for LR, SVM, LR+SVM, and the proposed method are 10, 9, 15, and 21 respectively. The Positive Accuracy of the proposed method (0.31) is twice that of the LR-based method's result (0.15). For the flare at level three, the proposed method is able

Table 2 Experimental Results

Level		Observation Positive				Observation Negative			
		LR	SVM	LR+SVM	AdaBoost	LR	SVM	LR+SVM	AdaBoost
0	Prediction Positive	52	46	45	46	28	16	14	14
	Prediction Negative	11	17	18	17	139	151	153	153
1	Prediction Positive	17	12	9	11	7	16	8	19
	Prediction Negative	48	53	56	54	158	149	157	146
2	Prediction Positive	10	9	15	21	2	14	27	26
	Prediction Negative	58	59	53	47	160	148	135	136
3	Prediction Positive	1	7	7	13	0	12	9	7
	Prediction Negative	33	27	27	21	196	184	187	189

Table 3 Comparison of All the Methods

Level		Correctness	True Positive	True Negative	Weighted True Rate	Positive Accuracy	Negative Accuracy	Weighted Accuarcy
0	LR	0.83	0.65	0.93	0.85	0.83	0.83	0.83
	SVM	0.85	0.74	0.89	0.85	0.73	0.9	0.85
	LR+SVM	0.86	0.76	0.89	0.86	0.72	0.92	0.86
	AdaBoost	0.87	0.77	0.9	0.86	0.73	0.92	0.87
1	LR	0.76	0.71	0.77	0.75	0.26	0.96	0.76
	SVM	0.7	0.43	0.74	0.65	0.18	0.9	0.7
	LR+SVM	0.72	0.53	0.74	0.67	0.14	0.95	0.72
	AdaBoost	0.68	0.37	0.73	0.63	0.17	0.88	0.68
2	LR	0.74	0.83	0.73	0.76	0.15	0.93	0.74
	SVM	0.68	0.39	0.72	0.62	0.13	0.93	0.68
	LR+SVM	0.65	0.35	0.72	0.61	0.22	0.83	0.65
	AdaBoost	0.68	0.45	0.74	0.66	0.31	0.84	0.68
3	LR	0.86	1	0.86	0.88	0.03	1	0.86
	SVM	0.83	0.37	0.87	0.8	0.21	0.94	0.83
	LR+SVM	0.84	0.44	0.87	0.81	0.21	0.95	0.84
	AdaBoost	0.88	0.65	0.9	0.86	0.38	0.96	0.88

to correctly predict 13 of the 34 samples, but the results of the other methods are only 1, 7, and 7 respectively. As a result, we may conclude that the LR-based method is almost ineffective for forecasting the occurrences of large flares, while the proposed method reveals the best performance for this task. Besides, the parameters b and c in Table 1 are also important aspects in evaluation of all the methods. Some resources are needed to perform the protective behavior. If the expected flare does not occur, the resources are wasted, and they can be regarded as a loss. However, the eruption of an unexpectedly large flare may cause enormous damage to a satellite or spacecraft that does not perform protective measures. On the other hand, if the flare is over-predicted, there will also be interruptions in satellite operation due to more times for the equipment being off. From this point of view, the proposed method is also superior to the other methods. As illustrated in Table 2 we can find that the parameters b and c of the proposed method are 7 and 21, which are both smaller than SVM and LR+SVM methods' results.

5 CONCLUSIONS

In this paper, we have proposed a flare prediction approach based on the AdaBoost algorithm. Experimental results on 230 active regions extracted from SOHO/MDI magnetograms show the effectiveness of the proposed method to predict different levels of flares. Based on the experimental results, we may find that none of the test methods can obtain the best results in all situations. The

LR based method performs better for the prediction of small levels of flares, while the proposed method achieves better results in predicting the occurrences of large flares. In the future, we will try to improve the AdaBoost algorithm to predict low level flares, and derive several novel predictors from the active regions for flare prediction.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 41174165 and 60973157), the Innovation Program Award for Graduate Students in Jiangsu Province, China (Nos. CXZZ12_0510 and CXZZ11_0625), and the Nanjing University of Information Science and Technology Excellent Teachers Scholarship for Overseas Studies for Ding.

References

Cui, Y., Li, R., Zhang, L., He, Y., & Wang, H. 2006, Sol. Phys., 237, 45
Dauphin, C., Vilmer, N., & Anastasiadis, A. 2007, A&A, 468, 273
Du, Z.-L., & Wang, H.-N. 2012, RAA (Research in Astronomy and Astrophysics), 12, 400
Fang, C. 2011, RAA (Research in Astronomy and Astrophysics), 11, 1377
Fang, C., Chen, P.-F., Jiang, R.-L., & Tang, Y.-H. 2010, RAA (Research in Astronomy and Astrophysics), 10, 83
Forbes, T. G., Linker, J. A., Chen, J., et al. 2006, Space Sci. Rev., 123, 251
Freund, Y., Schapire, R. E., 1997, J. Comput. Syst. Sci., 1, 119
Gan, W. Q., Rieger, E., & Fang, C. 1993, ApJ, 416, 886
Jing, J., Song, H., Abramenko, V., Tan, C., & Wang, H. 2006, ApJ, 644, 1273
Li, R., Wang, H.-N., He, H., Cui, Y.-M., & Du, Z.-L. 2007, ChJAA (Chin. J. Astron. Astrophys.), 7, 441
Li, Y. P., & Gan, W. Q. 2011, Sol. Phys., 269, 59
Qahwaji, R., & Colak, T. 2007, Sol. Phys., 241, 195
Sakurai, T. 1976, PASJ, 28, 177
Schapire, R. E. 2003, Nonlinear Estimation and Classification, 171 of Lecture Notes in Statistics (Springer)
Song, H., Tan, C., Jing, J., et al. 2009, Sol. Phys., 254, 101
Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, Sol. Phys., 255, 91
Yu, D., Huang, X., Hu, Q., et al. 2010a, ApJ, 709, 321
Yu, D., Huang, X., Wang, H., et al. 2010b, ApJ, 710, 869
Yuan, Y., Shih, F. Y., Jing, J., & Wang, H.-M. 2010, RAA (Research in Astronomy and Astrophysics), 10, 785
Zhang, J., Wang, J., Deng, Y., & Wu, D. 2001, ApJ, 548, L99