Research in Astron. Astrophys. **2010** Vol. **10** No. **8**, 785–796
http://www.raa-journal.org *http://www.iop.org/journals/raa*

Research in Astronomy and Astrophysics

Automated flare forecasting using a statistical learning technique

Yuan Yuan[1], Frank Y. Shih[1], Ju Jing[2] and Hai-Min Wang[2]

[1] Computer Vision Laboratory, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA; *yy46@njit.edu*
[2] Space Weather Research Laboratory, Physics Department, New Jersey Institute of Technology, Newark, NJ 07102, USA

Received 2009 March 19; accepted 2010 April 10

Abstract We present a new method for automatically forecasting the occurrence of solar flares based on photospheric magnetic measurements. The method is a cascading combination of an ordinal logistic regression model and a support vector machine classifier. The predictive variables are three photospheric magnetic parameters, i.e., the total unsigned magnetic flux, length of the strong-gradient magnetic polarity inversion line, and total magnetic energy dissipation. The output is true or false for the occurrence of a certain level of flares within 24 hours. Experimental results, from a sample of 230 active regions between 1996 and 2005, show the accuracies of a 24-hour flare forecast to be 0.86, 0.72, 0.65 and 0.84 respectively for the four different levels. Comparison shows an improvement in the accuracy of X-class flare forecasting.

Key words: Sun: flares — Sun: magnetic fields

1 INTRODUCTION

The sudden and intense release of energy stored in solar magnetic fields generates solar flares (Dauphin et al. 2007), which can have a significant impact on the near earth space environment (so called space weather). The development of fully automatic programs to detect (e.g. Qu et al. 2003, 2004) and forecast flares is regarded as one of the tasks to process the large amount of data in this field accurately and efficiently.

At present, a number of different flare forecasting approaches and systems have been developed based on photospheric magnetic field observations or sunspot-group characteristics. For instance, "Theophrastus," a system developed by the Space Weather Prediction Center of NOAA, is mainly based on the correlation between solar flare production and sunspot-group classification (McIntosh 1990). At Big Bear Solar Observatory, Gallagher et al. (2002) used the historical average of flare numbers by the McIntosh classification to develop a solar flare prediction system which estimated the probabilities for each active region to produce C-, M-, or X-class flares. Barnes et al. (2007) adopted discriminant analysis to accomplish solar flare forecasting within 24 hours using a large combination of vector magnetic field measurements obtained by the University of Hawaii Imaging Vector Magnetograph. Li et al. (2008) proposed a solar flare forecasting method based on support vector machines in which the sunspot area, the sunspot magnetic class, the McIntosh class of the sunspot group and the 10 cm solar radio flux were chosen as precursors. Georgoulis & Rust (2007) defined a new measurement called the effective connected magnetic field, and their experimental

results, based on 298 active regions during a 10 year period of solar cycle 23, showed that this measure was an efficient flare-forecasting criterion. Qahwaji & Colak (2007) put forward a short-term solar flare prediction method using machine learning and sunspot associations, in which the authors had compared the performance of the proposed method with two other machine learning algorithms.

Different from the approaches mentioned above, Wheatland (2005) designed a Bayesian approach to solar flare prediction in which only the event statistics of flares already observed was used as predictors, however this approach has not been tested on a large data set.

In this paper, we present a new method for the automatic forecasting of the occurrence of solar flares over 24 hours following the time when a magnetogram is recorded. Our method is a continuation and extension of the method proposed by Song et al. (2009), which has some limitations in forecasting X-class flares, as it has to use an arbitrarily imposed threshold. Our method is split into two cascading steps. In the first step, the logistic regression model is used to map three magnetic parameters of each active region into four probabilities; the support vector machine classifier is then utilized to map the four probabilities onto a binary label which is the final output. Experimental results illustrate how the proposed method performs better when it comes to X-class flare forecasting.

The paper is organized as follows. The definitions of the predictive variables (i.e., three magnetic parameters) used in this study are introduced in Section 2. The proposed flare forecasting method is described in Section 3. Experimental results are shown in Section 4. Finally, a conclusion is drawn in Section 5.

2 DATA DESCRIPTION

2.1 Predictive Variables

To be consistent with the work of Song et al. (2009), the same predictive variables are used. The predictive variables of Song et al. (2009) are composed of

1. Total unsigned magnetic flux, $T_{\rm flux}$, which is the integration of pixel intensity over the area of an active region,
$$T_{\rm flux} = \iint |B_z| dxdy, \tag{1}$$
where B_z is the pixel intensity of MDI magnetograms.
2. Length of the strong-gradient magnetic polarity inversion line, $L_{\rm gpi}$, which was first studied by Falconer et al. (2003) as a measure to predict coronal mass ejections. Jing et al. (2006) illustrated the correlation between $L_{\rm gpi}$ and flare productivity of active regions. As illustrated in Song et al. (2009), $L_{\rm gpi}$ is the total number of pixels on which the gradient $|\nabla_\perp B_z|$ is greater than a threshold, which is 50 G Mm^{-1} as chosen by Song et al. (2009). The definition of $|\nabla_\perp B_z|$ is as follows (Song et al. 2009):
$$|\nabla_\perp B_z| = \left[\left(\frac{dB_z}{dx} \right)^2 + \left(\frac{dB_z}{dy} \right)^2 \right]^{1/2}. \tag{2}$$
3. Total magnetic energy dissipation, $E_{\rm diss}$, proposed by Abramenko et al. (2003), was also studied by (Jing et al. 2006; Song et al. 2009) in exploring its correlation between flare productivity of active regions. According to Abramenko et al. (2003),
$$E_{\rm diss} = \iint 4 \left[\left(\frac{dB_z}{dx} \right)^2 + \left(\frac{dB_z}{dy} \right)^2 \right] + 2 \left(\frac{dB_z}{dx} + \frac{dB_z}{dy} \right)^2 dxdy, \tag{3}$$
where the integration is performed over the area of an active region.

Table 1 Mean Value and Standard Deviation of Predictive Parameters

Active Region	Number of Active	$L_{\rm gpi}$ (Mm)		$T_{\rm flux}$ (10^{22}Mx)		$E_{\rm diss}$ (10^8 erg cm^{-3})	
Level	Regions	Mean	Dev.	Mean	Dev.	Mean	Dev.
3	34	118.74	79.88	7.02	3.15	15.38	7.76
2	68	64.28	46.79	5.03	2.72	10.58	5.59
1	65	62.12	46.61	4.95	2.86	10.47	5.88
0	63	10.84	15.19	1.72	1.19	3.67	2.58

We chose these parameters mainly because: (1) all three can be derived from the line-of-sight magnetograms; and (2) all three moderately correlate with the flare productivity of active regions and show their forecasting utility in the previous study by Jing et al. (2006) and Song et al. (2009).

2.2 Data Collection

The three magnetic parameters introduced above were derived from the magnetograms produced by the Michelson Doppler Imager (MDI), which is an instrument onboard the Solar and Heliospheric Observatory (SOHO).

Our study uses the same dataset as was used by Song et al. (2009), which focuses on active regions between 1996 and 2005. It covers almost the entire solar cycle 23 which peaked in 2001. A total of 230 sample active regions were selected using the following criteria: (1) the center location of an active region is close to the solar disk center (within ± 40 degrees in longitude and ± 40 degrees in latitude); (2) the MDI full disk magnetograms are available; (3) since an active region may appear on the solar surface for a few days, it is treated as a different sample on different dates; (4) the first magnetogram of the 15 magnetograms taken by MDI each day is chosen.

2.3 Correlation between Magnetic Parameters and Flare Productivity

Using the same criteria as (Song et al. 2009), active regions are categorized into four levels according to the most powerful flare produced: an active region is classified as level 0 if it is flaring-quiet or only produces A and/or B class flares; an active region is classified as level 1 if it produces at least one C-class flare but no M- or X- class flares; Level 2 corresponds to those active regions which produce at least one M-class flare but no X-class flares; Level 3 corresponds to those active regions which produce at least one X class flare.

Based on the 230 active regions in our dataset (see Appendix), we can discern the correlations between magnetic parameters and flare productivity, which are summarized in Table 1.

From Table 1, we notice that the mean value of the length of the strong-gradient magnetic polarity inversion line of the 34 level 3 active regions is 118.74, which is much larger than that of the 68 level 2 active regions (64.28). The mean value of the length of the strong-gradient magnetic polarity inversion line of 68 level 2 active regions is 64.28, which is slightly larger than that of the 65 level 1 active regions (62.12). The mean value of the length of the strong-gradient magnetic polarity inversion line of 63 level 0 active regions is 10.84, which is much less than that of other levels of active regions. For total unsigned magnetic flux and total magnetic energy dissipation, the same kind of trend follows. However, since the fluctuations are large (almost half of the mean values), it is impossible to do precise flare forecasting based on those parameters.

Based on the correlations described above, we combine statistical and machine learning methods to perform flare forecasting.

3 FORECASTING METHOD

In previous studies, there are mainly two types of flare forecasting methods. The first type is based on pattern recognition, such as a Support Vector Machine-based (SVM-based) method (Li et al. 2008). During this kind of analysis, some predictive parameters of a given active region are extracted, and then the predictive parameters are fed into a trained classifier. The output of the classifier (usually a label indicating which class of flare is likely to occur) is the final forecasting result. The disadvantage of this type is that the output is only a label, which does not provide information on how much confidence can be placed on each forecast. For example (see Fig. 1), both sample A and sample B will be classified as the same class, but intuitively we should be more confident that B belongs to this class than A, since A is on the boundary. However, since the output is only a label, we do not get this crucial information.

The second type is based on probability analysis, such as ordinal logistic regression (Song et al. 2009). During this kind of analysis, some predictive parameters of a given active region are extracted, and then those predictive parameters are fed into a trained statistical model, and the output of the model is the probability that a flare event will occur. Of course, using a threshold value (generally 0.5), we can turn the probability into a binary forecast. However, it is not an easy job to choose a good threshold value, and the de facto standard threshold (0.5) is not always the best, as illustrated in Song et al. (2009), where the authors chose 0.25 as the threshold for X-class flare prediction.

In this paper, the proposed method is split into two steps (see Fig. 2). In the first step, we adopt ordinal logistic regression to map the input (three predictive parameters of a given active region) to four outputs (the probabilities of the given active region belonging to each of the four levels). Secondly, the four outputs are fed into a support vector machine; the output of the support vector machine tells us whether the given active region belongs to one level or not.

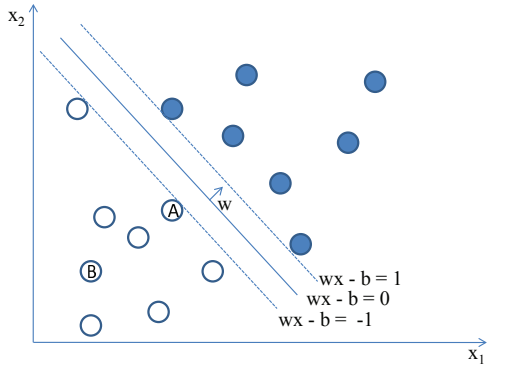


Fig. 1 An illustration of the support vector machine classifier.

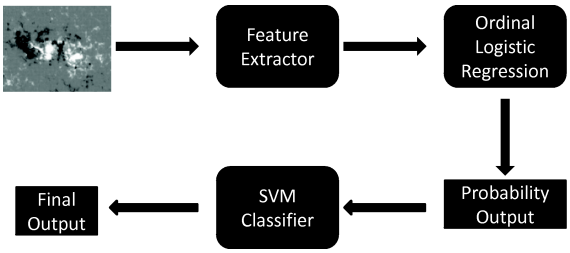


Fig. 2 The working flow of the proposed forecasting system.

Generally, the first step is enough for a flare forecasting system. The purpose of the second step is three fold. First, it is hard to assess the performance of the first step since the outputs are probabilities instead of a definite answer. Secondly, users sometimes want a definite answer instead of a probability. Thirdly, the outputs of the second step can be used to compare with other research whose outputs are only binary labels.

3.1 Probability Prediction Using Ordinal Logistic Regression

Used for Bernoulli-distributed dependent variables, logistic regression is a generalized linear model that uses the logit as its link function (McCullagh & Nelder 1989). One common application of logistic regression is to estimate the probability of the occurrence of an event from predictive variables. Ordinal logistic regression is used to map predictive variables into probabilities of the occurrence of flares by Song et al. (2009). The comparison made by Song et al. (2009) shows that their forecasting results are better than those of the Solar Data Analysis Center and NOAA's Space Weather Prediction Center, which illustrates the usefulness of ordinal logistic regression in flaring probability estimation.

Suppose that the data in a dataset belong to L categories and $P(D = g)$ is the probability that an event which belongs to category g would occur given predictive variables $\mathbf{X}$, then, according to Kleinbaum & Klein (2002),

$$\begin{aligned} &P(D = g) = P(D \geq g) - P(D \geq g + 1), \\ &P(D \geq g) = \frac{1}{1 + e^{-(\alpha_g + \beta^{\mathbf{T}}\mathbf{X})}}, \\ &g = 1, 2, 3, ..., L. \end{aligned} \tag{4}$$

Given a training dataset composed of predictive variables and response category pairs, the parameters $\alpha_g, g = 1, 2, 3, .., L$ and β in the above equation can be calculated using the method of estimation called maximum likelihood (Hosmer & Lemeshow 2000).

The application of ordinal logistic regression to flare forecasting is as follows:

1. Training: The training data contain several samples; each sample is composed of three photospheric magnetic features of an active region and the level of the given active region.
2. Forecasting: Using the ordinal logistic regression model, for a given active region, at first, we figure out its three photospheric magnetic features, and then feed these three variables into the model. The output of the model contains four elements, which correspond to the probabilities that the given active region belongs to level 0, 1, 2, or 3.

3.2 Binary Forecasting Using Support Vector Machines

An SVM is a supervised learning method used for classification (Boser et al. 1992), whose principle is to minimize the structural risk (Vapnik 1995). An SVM tries to find a plane in an n-dimensional space that separates input data into two classes. The larger the distance from the plane to the two different classes of data points in the n-dimensional space, the smaller the classification error (Cortes & Vapnik 1995).

Given training vectors $\mathbf{x}_i \in R^d, i = 1, 2, ..., n$ in two classes labeled by a vector $\mathbf{y} \in R^n$ where $y_i = \{-1, 1\}, i = 1, 2, ..., n$. The training of a support vector machine is equivalent to solving the following optimization problem (Fan et al. 2005):

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \left(\frac{1}{2} \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \right), \\ &\mathbf{y}^T \boldsymbol{\alpha} = 0, \\ &0 \leq \alpha_i \leq C, i = 1, 2, ..., n \end{aligned} \tag{5}$$

Table 2 A Sample Contingency Table

	Observation Positive	Observation Negative
Forecasting Positive	a	b
Forecasting Negative	c	d

where $\mathbf{e}$ is the vector of all ones, $C > 0$ is the upper bound, $\mathbf{Q}$ is an n by n positive semi-definite matrix, $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, and $K(\mathbf{x}_i, \mathbf{x}_j) \equiv (\gamma \mathbf{x}_i^T \mathbf{y}_j + \zeta)^d$ is the kernel function. The decision function is:

$$f(\mathbf{x}) = \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{6}$$

The prediction of any test data $\mathbf{x}$ is $\mathrm{sgn}\,(f(\mathbf{x})) \in \{-1, 1\}$.

For flare forecasting, the training and forecasting procedures of a support vector machine are as follows:

1. Training: The training data contain several samples; each sample is composed of four probabilities (the output of ordinal logistic regression) and one label (–1 or 1). If a given active region indeed belongs to one level, the label is 1; otherwise, the label is –1.
2. Forecasting: Given an active region, at first, we figure out its three photospheric magnetic features. Then we feed these three variables into the ordinal logistic model to generate the output which contains four probabilities. Finally, we feed the four probabilities into the support vector machine trained above. If the output of the support vector machine is 1, the estimation is that the given active region belongs to one level; otherwise, it does not.

4 EXPERIMENTAL RESULTS

The proposed flare forecasting method is implemented in MATLAB (Moler 2004), which contains a procedure to fit a logistic regression model. The implementation also utilizes LIBSVM (Chang & Lin 2001), which is a software package for support vector classification. The parameters adopted for LIBSVM are as follows: nu-Support Vector Classification of polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) \equiv (0.01 \mathbf{x}_i^T \mathbf{y}_j)^3$.

We use four different trained SVM classifiers to perform yes/no forecasting for four different levels. The outputs of the first step (four probabilities) and the corresponding labels are sent to the four SVM classifiers to train them in the second step. The training procedures are almost the same for the four SVM classifiers except that different labels are used, i.e., when training a level-n SVM classifier, the four probabilities and a label which indicates whether the given sample belongs to level-n are fed into the SVM classifier, where $n = 0$, 1, 2 or 3. Alternatively, we can use a multi-class SVM classifier. In that way, only one multi-class SVM classifier is needed instead of four different binary SVM classifiers. We plan to implement the multi-class SVM classifier version of the forecasting method in the future.

Leave-one-out cross-validation is used to assess the prediction performance. For 230 samples, during each test case, 229 samples are used for training, and the remaining one is used for testing. If the predicted result is the same as the observation, it is positive; otherwise, it is negative. The process is repeated 230 times. Different samples are used for training and testing each time.

To assess the performance of the proposed method, seven measurements are used, which are correctness, true positive, true negative, weighted true rate, positive accuracy, negative accuracy, and weighted accuracy. All these seven measurements can be derived from the contingency table of the experiment. For a given contingency table like Table 2, we can derive the seven measurements as follows:

1. correctness $= (a+d)/(a+b+c+d)$;
2. true positive $= a/(a+b)$;
3. true negative $= d/(c+d)$;
4. weighted true rate $= a/(a+b)*(a+c)/(a+b+c+d)+d/(c+d)*(b+d)/(a+b+c+d)$;
5. positive accuracy $= a/(a+c)$;
6. negative accuracy$= d/(b+d)$;
7. weighted accuracy $= a/(a+c)*(a+c)/(a+b+c+d)+d/(b+d)*(b+d)/(a+b+c+d)$.

To compare the performance of the proposed method with the Logistic-Regression-based method (Song et al. 2009) and SVM-based method (Li et al. 2008), we do the experiments on the same dataset and the experimental results are illustrated in Figures 3, 4, 5 and 6. These four figures contain not only the contingency tables of each experiment, but also bar charts to illustrate the seven measures derived from contingency tables to help us compare the performances of the three different flare forecasting methods. Please note, among the seven measures, positive accuracy is the most important measure in flare forecasting in that a miss (forecasting no flare, but flares occur) is worse than a false alarm (forecasting the occurrence of a flare, but it does not happen). The higher the value of positive accuracy, the less events are missed.

	Observation Positive	Observation Negative
Forecasting Positive	52	28
Forecasting Negative	11	139

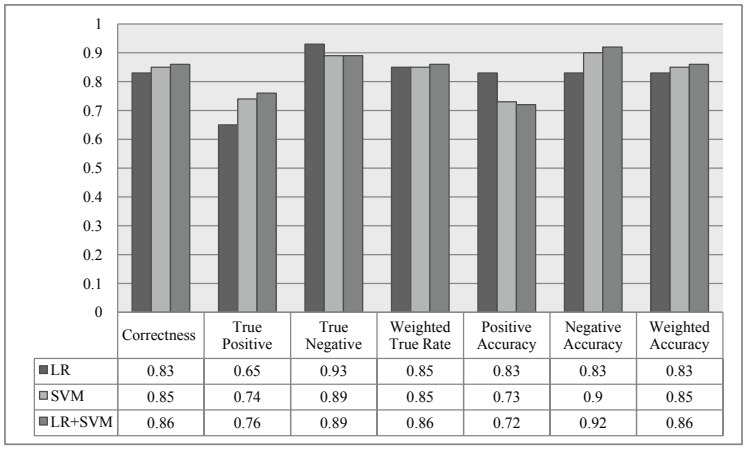
(a) Contingency table of logistic-regression-based method

	Observation Positive	Observation Negative
Forecasting Positive	46	16
Forecasting Negative	17	151

(b) Contingency table of SVM-based method

	Observation Positive	Observation Negative
Forecasting Positive	45	14
Forecasting Negative	18	153

(c) Contingency table of the proposed method



	Correctness	True Positive	True Negative	Weighted True Rate	Positive Accuracy	Negative Accuracy	Weighted Accuracy
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

(d) Comparison of methods

Fig. 3 Experiment on level zero.

	Observation Positive	Observation Negative
Forecasting Positive	17	7
Forecasting Negative	48	158

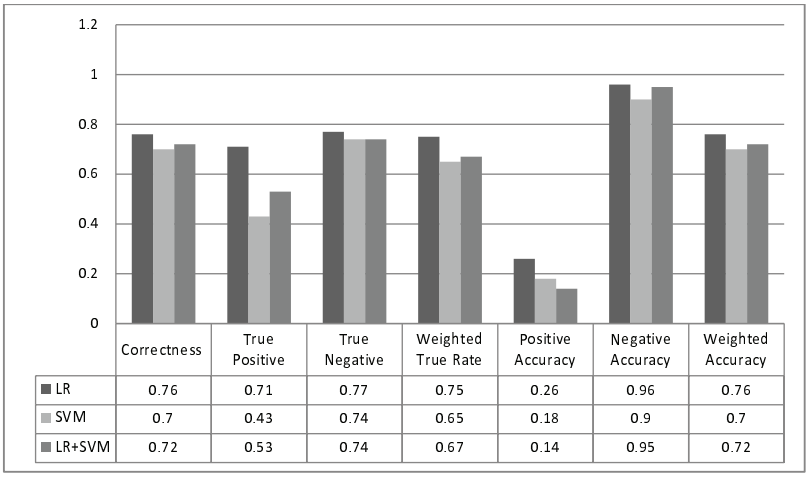
(a) Contingency table of logistic-regression-based method

	Observation Positive	Observation Negative
Forecasting Positive	12	16
Forecasting Negative	53	149

(b) Contingency table of SVM-based method

	Observation Positive	Observation Negative
Forecasting Positive	9	8
Forecasting Negative	56	157

(c) Contingency table of the proposed method



	Correctness	True Positive	True Negative	Weighted True Rate	Positive Accuracy	Negative Accuracy	Weighted Accuracy
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

(d) Comparison of methods

Fig. 4 Experiment on level one.

Figures 3, 4, 5 and 6 show the forecasting results for levels zero, one, two and three respectively, e.g., for level zero forecasting, all these 230 active regions in our dataset are classified into two groups according to whether they belong to level zero, and then the forecasting models are trained, and then tested.

Predicting the occurrence of X-class flares is the most important task of flare forecasting. As we can see from panel (a) in Figure 6, the Logistic-Regression-based method does not work well for forecasting X-class flares. Only 1 of the 34 X-class flares is forecasted correctly. At the same time, the SVM-based method and our proposed method can correctly forecast 7 of the 34 X-class flares, which is an improvement over the Logistic-Regression-based method. From Figure 5, we also notice that our proposed method outperforms the other two methods on level two (M-class flares) forecasting.

The experimental results also show that our proposed flare forecasting method outperforms the SVM-based method on level one and level three forecasting. However, our proposed method is surpassed by the SVM-based method on level two forecasting, but the difference is tiny. The performances of these two methods on level zero forecasting are almost the same.

	Observation Positive	Observation Negative
Forecasting Positive	10	2
Forecasting Negative	58	160

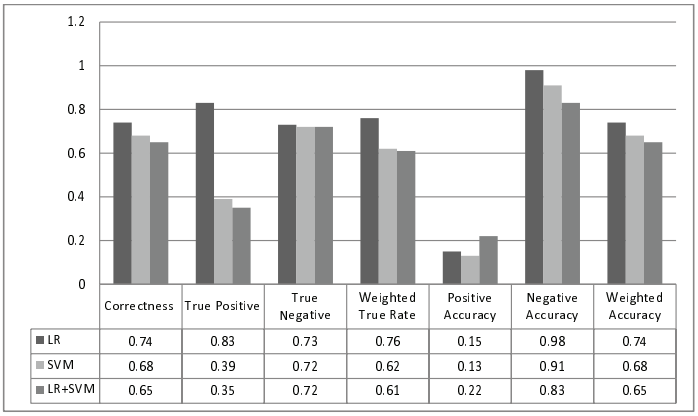
(a) Contingency table of logistic-regression-based method

	Observation Positive	Observation Negative
Forecasting Positive	9	14
Forecasting Negative	59	148

(b) Contingency table of SVM-based method

	Observation Positive	Observation Negative
Forecasting Positive	15	27
Forecasting Negative	53	135

(c) Contingency table of the proposed method



	Correctness	True Positive	True Negative	Weighted True Rate	Positive Accuracy	Negative Accuracy	Weighted Accuracy
LR	0.74	0.83	0.73	0.76	0.15	0.98	0.74
SVM	0.68	0.39	0.72	0.62	0.13	0.91	0.68
LR+SVM	0.65	0.35	0.72	0.61	0.22	0.83	0.65

(d) Comparison of methods

Fig. 5 Experiment on level two.

5 CONCLUSIONS

In this paper, we propose a solar flare prediction method based on ordinal logistic regression and a support vector machine. For 230 active regions between 1996 and 2005, their magnetic parameters ($L_{\rm gpi}, T_{\rm Flux}, E_{\rm diss}$) measured from SOHO MDI magnetograms are extracted and used for training. Our results can be summarized as follows:

1. The proposed method is a valid flare forecasting method, which performs almost equally well with the SVM-based method.
2. Although comparison shows that the positive accuracy of the proposed method is better than that of the Logistic-Regression-based method on X-class flare forecasting, the true positive rate (0.44) and positive accuracy (0.21) are still very low, which means we may fail to predict some occurrences of the X-class flares.
3. Since the proposed method is split into two cascading steps, one extra advantage of the proposed method over the SVM-based method is that we know the confidence of the forecasting results. For example, when both of these two methods classify one active region into level three, we can derive the confidence level by examining the output of the first step. Since the output of the first step (the output of logistic regression) contains four probabilities (the four probabilities that a given active region belongs to the four levels), the higher the fourth probability, the more confidence we can have about the forecast results of X-class flares (corresponding to level three).

	Observation Positive	Observation Negative
Forecasting Positive	1	0
Forecasting Negative	33	196

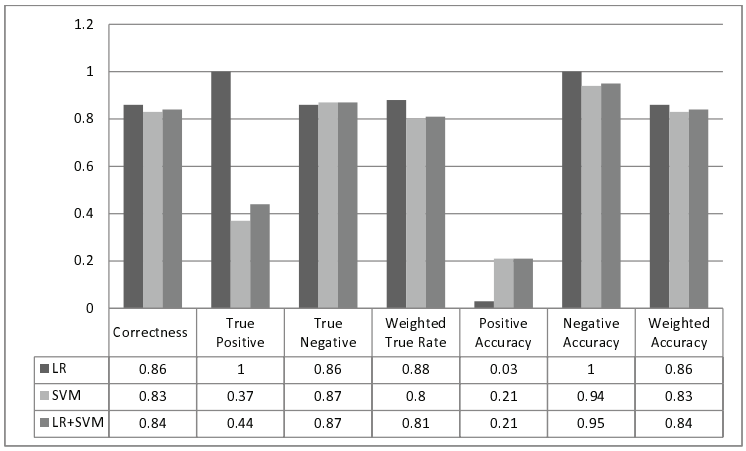
(a) Contingency table of logistic-regression-based method

	Observation Positive	Observation Negative
Forecasting Positive	7	12
Forecasting Negative	27	184

(b) Contingency table of SVM-based method

	Observation Positive	Observation Negative
Forecasting Positive	7	9
Forecasting Negative	27	187

(c) Contingency table of the proposed method



	Correctness	True Positive	True Negative	Weighted True Rate	Positive Accuracy	Negative Accuracy	Weighted Accuracy
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

(d) Comparison of methods

Fig. 6 Experiment on level three.

So far, our prediction model is limited to those magnetic parameters obtained only through SOHO MDI magnetograms. There are several other physical parameters (such as magnetic free-energy, electric current and helicity injections) that we are currently investigating, and from which we anticipate that the performance of our method can be improved. Similar to some other machine learning techniques, our method is scalable with regard to accepting new parameters. In the future, after deriving several new magnetic parameters from vector magnetograms from the Solar Dynamic Observatory and *Hinode*, the new values should help us to improve the performance of the proposed forecasting method. In addition, we plan to incorporate measures such as sunspot structure change (Chen et al. 2007) and topology of solar magnetic fields (Zhao et al. 2008) to improve the performance of the proposed forecasting method.

Acknowledgements This work is supported by NSF under grants ATM-0716950, ATM-0745744 and NASA under grant NNXO-8AQ90G. We thank the referee for providing constructive comments and help. We also thank Dr. Louis Lanzerotti for his valuable comments.

Appendix A: HISTOGRAMS OF THE DATA SET USED IN THE EXPERIMENT

Figures A.1, A.2 and A.3 illustrate the histograms of the length of the strong gradient inversion line, total unsigned flux and energy dissipation. Please note the values are scaled to 0 and 1, and the unit is shown below each graph. The height of a bar denotes the number of samples whose corresponding parameters are within some range. Within each range, different colored bars are used to differentiate the samples into different levels.

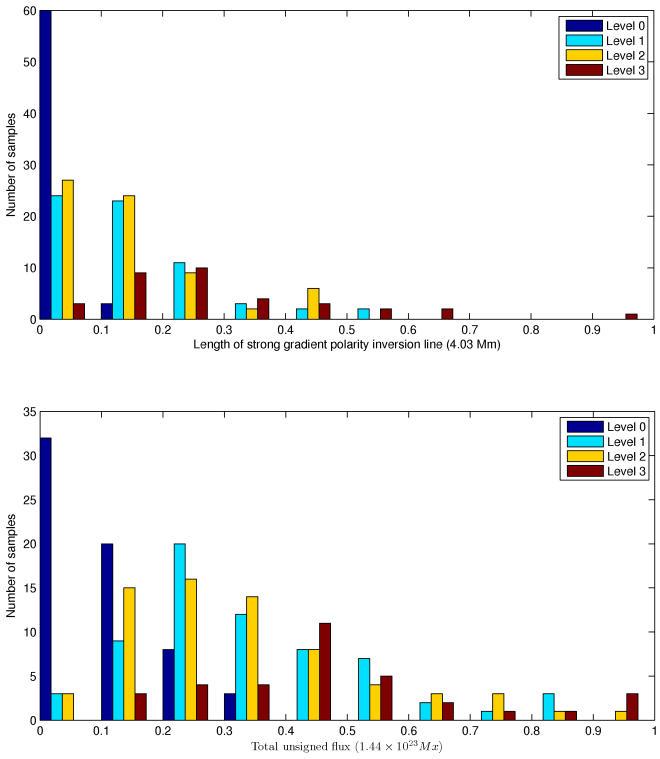


Fig. A.1 Histogram of the second parameter for the four different levels.

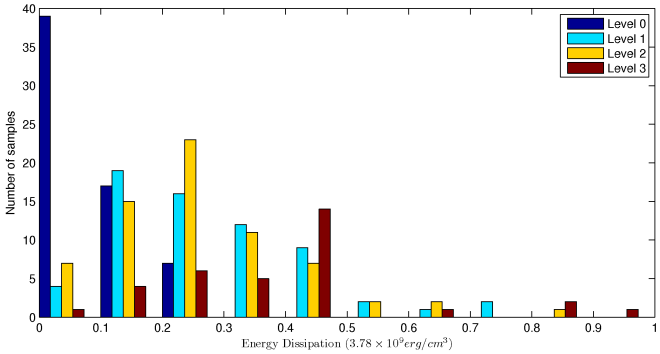


Fig. A.2 Histogram of the third parameter for the four different levels.

For example, the height of the blue bar in Figure A.3 within range 0 and 0.1 is 39, which means there are 39 level 0 samples whose energy dissipation is within the range 0 and 3.78×10^8 erg cm^{-3}. As we can see, the blue bars (which correspond to level 0 samples) are mainly distributed in the lower ranges, and their heights decrease as the values increase. The red bars (which correspond to level 3 samples) can reach higher ranges, which coincide with our observations that samples with higher values of these parameters are more likely to produce X-class flares.

References

Abramenko, V. I., Yurchyshyn, V. B., Wang, H., Spirock, T. J., & Goode, P. R. 2003, ApJ, 597, 1135

Barnes, G., Leka, K. D., Schumer, E. A., & Della-Rose, D. J. 2007, Space Weather, 5, 9002

Boser, B., Guyon, I., & Vapnik, V. 1992, A training algorithm for optimal margin classifiers, in the Fifth Annual Workshop on Computational Learning Theory, 144

Chang, C.-C., & Lin, C.-J. 2001, LIBSVM: a library for support vector machines, software available at *http://www.csie.ntu.edu.tw/ cjlin/libsvm*

Chen, W.-Z., Liu, C., Song, H., et al. 2007, ChJAA (Chin. J. Astron. Astrophys.), 7, 733

Cortes, C., & Vapnik, V. N. 1995, Support-vector network, Machine Learning, 20, 273

Dauphin, C., Vilmer, N., & Anastasiadis, A. 2007, A&A, 468, 273

Falconer, D. A., Moore, R. L., & Gary, G. A. 2003, J. Geophys. Res. (Space Physics), 108, 1380

Fan, R. E., Chen, P. H., & Lin, C. J. 2005, Journal of Machine Learning Research, 6, 1889

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

Georgoulis, M. K., & Rust, D. M. 2007, ApJ, 661, L109

Hosmer, D. W., & Lemeshow, S. 2000, Applied Logistic Regression (second ed.; Wiley-Interscience Publication)

Jing, J., Song, H., Abramenko, V., Tan, C., & Wang, H. 2006, ApJ, 644, 1273

Kleinbaum, D. G., & Klein, M. 2002, Logistic Regression: A Self-Learning Text (second ed.; Springer)

Li, R., Cui, Y., He, H., & Wang, H. 2008, Advances in Space Research, 42, 1469

McCullagh, P., & Nelder, J. A. 1989, Generalized Linear Models (second ed.; Chapman and Hall/CRC)

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

Moler, C. B. 2004, Numerical Computing with Matlab, Society for Industrial Mathematics

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

Qu, M., Shih, F. Y., Jing, J., & Wang, H. 2003, Sol. Phys., 217, 157

Qu, M., Shih, F., Jing, J., & Wang, H. 2004, Sol. Phys., 222, 137

Song, H., Tan, C., Jing, J., Wang, H., Yurchyshyn, V., & Abramenko, V. 2009, Sol. Phys., 254, 101

Vapnik, V. N. 1995, The Nature of Statistical Learning Theory (Springer)

Wheatland, M. S. 2005, Space Weather, 3, 7003

Zhao, H., Wang, J.-X., Zhang, J., et al. 2008, ChJAA (Chin. J. Astron. Astrophys.), 8, 133