The Astrophysical Journal, 709:321–326, 2010 January 20 doi:10.1088/0004-637X/709/1/321


# SHORT-TERM SOLAR FLARE PREDICTION USING MULTIRESOLUTION PREDICTORS

Daren Yu[1], Xin Huang[1], Qinghua Hu[1], Rui Zhou[1], Huaning Wang[2], and Yanmei Cui[3]
[1] Harbin Institute of Technology, No. 92 West Da-Zhi Street, Harbin, Heilongjiang Province, China; huangxinhit@yahoo.com.cn
[2] National Astronomical Observatories, 20A Datun Road, Chaoyang District, Beijing, China
[3] Center for Space Science and Applied Research, No. 1 Nanertiao, Zhongguancun, Haidian District, Beijing, China
*Received 2009 May 12; accepted 2009 December 2; published 2009 December 30*

## ABSTRACT

Multiresolution predictors of solar flares are constructed by a wavelet transform and sequential feature extraction method. Three predictors—the maximum horizontal gradient, the length of neutral line, and the number of singular points—are extracted from *Solar and Heliospheric Observatory*/Michelson Doppler Imager longitudinal magnetograms. A maximal overlap discrete wavelet transform is used to decompose the sequence of predictors into four frequency bands. In each band, four sequential features—the maximum, the mean, the standard deviation, and the root mean square—are extracted. The multiresolution predictors in the low-frequency band reflect trends in the evolution of newly emerging fluxes. The multiresolution predictors in the high-frequency band reflect the changing rates in emerging flux regions. The variation of emerging fluxes is decoupled by wavelet transform in different frequency bands. The information amount of these multiresolution predictors is evaluated by the information gain ratio. It is found that the multiresolution predictors in the lowest and highest frequency bands contain the most information. Based on these predictors, a C4.5 decision tree algorithm is used to build the short-term solar flare prediction model. It is found that the performance of the short-term solar flare prediction model based on the multiresolution predictors is greatly improved.

*Key words:* methods: statistical – Sun: activity – Sun: flares – Sun: magnetic fields – Sun: photosphere

*Online-only material:* color figures

## 1. INTRODUCTION

Many efforts have been made to find powerful predictors of solar flares. McIntosh (1990) defined the McIntosh classification to represent morphological characteristics of active regions. The McIntosh parameters act as proxies for the magnetic properties of an active region. The Mount Wilson classification was an early magnetic classification of sunspots. Sammis et al. (2000) confirmed the relation between magnetic classification $\delta$ and large flares; furthermore, they confirmed that the region classified $\beta\gamma\delta$ produces many more large flares than other regions of comparable size. Falconer (1997) found that the total X-ray brightness of an entire active region is correlated with the total length of neutral lines on which the magnetic field is both strong and strongly sheared in the same active region. The magnetic field of the active region is not only related to the flares but also correlated with the coronal mass ejections (Moore et al. 2001; Falconer et al. 2006). Leka & Barnes (2003) extracted many parameters from the magnetic vector field to identify whether flares should occur. The power spectra of the line-of-sight magnetograms along with its correlation with flare productivity were discussed by Abramenko (2005). McAteer et al. (2005) quantified the magnetic complexity of active regions using a fractal dimension measure. Jing et al. (2006) proposed three parameters—the mean value of spatial magnetic gradients at strong-gradient magnetic neutral lines, the length of strong-gradient magnetic neutral lines, and the total magnetic energy dissipated in a layer of 1 m during 1 s over the active region's area—from Michelson Doppler Imager (MDI) magnetograms. The magnitude scaling correlations between these parameters and the flare productivity of active regions are explored. Cui et al. (2006) proposed three parameters—the maximum horizontal gradient, the length of the neutral line, and the number of singular points—to describe the nonpotentiality and complexity of the photospheric magnetic field in active regions. Cui et al. (2007) analyzed the statistical relationships among solar flares, magnetic gradient, and magnetic shear extracted from vector magnetograms of the Huairou Solar Observing Station. These predictors have been taken into account to build a solar flare prediction model (He et al. 2008). Georgoulis & Rust (2007) defined the effective connected magnetic field ($B_{\rm eff}$) as a single metric of the flaring potential in active regions. Schrijver (2007) pointed out that solar flares result from electromagnetic instability within strong magnetic field regions, and the total unsigned flux $R$ within ~15 Mm of strong-field, high-gradient polarity separation lines was used to forecast flares.

Based on various types of predictors, many solar flare prediction models have been built. Miller (1989) developed an expert system (WOLF) to analyze the solar active region and predict the probable occurrence of solar flares. McIntosh (1990) prompted use of the McIntosh classification of sunspots in an expert system (Theo) for predicting X-ray flares, and McIntosh classifications were considered a guide in predicting solar flares at the Space Environment Center of the National Oceanic and Atmospheric Administration (Bornmann & Shaw 1994); they were also used to provide the initial flaring probability for the system at http://www.solarmonitor.org (Gallagher et al. 2002). Bradshaw et al. (1989) constructed a three-layer back-propagation neural network to forecast flares using the McIntosh classification. Based on measures of the solar magnetic field, Wang et al. (2008) introduced a solar flare forecasting model supported with an artificial neural network. Li et al. (2007) presented a method combining the support vector machine and the $k$-nearest neighbors to construct a solar flare forecasting model. Qahwaji & Colak (2007) presented a hybrid system that combines a support vector machine and a cascade-correlation neural network for automatic short-term solar flare prediction. Discriminant analysis was applied to numerous photospheric

magnetic parameters of active regions to produce a binary categorization of a region as flaring or non-flaring by Leka & Barnes (2007). This approach was extended to a probability forecast in Barnes et al. (2007). Wheatland (2004) pointed out that the past history of occurrence of flares is an important indicator of future flare production, and a Bayesian approach to flare prediction using the flaring records of an active region was proposed.

Solar activity is a complex physical phenomenon, requiring research with different time or size scales. There are some studies on the size scales of magnetic fields of active regions. Portier-Fozzani et al. (2001) applied a Multiscale Vision Model based on a shift-invariant discrete wavelet algorithm to analyze *Solar and Heliospheric Observatory*/Extreme Ultraviolet Imaging Telescope (*SOHO*/EIT) images. Delouille et al. (2005) showed three applications of the continuous wavelet spectrum in the analysis of *SOHO*/EIT images. In Hewett et al. (2008), the continuous wavelet technique was used to investigate the multiscale structure of an active region using magnetograms obtained by *SOHO*/MDI. In Ireland et al. (2008), two different multiresolution analyses were used to decompose the structure of active region magnetic flux into concentrations of different size scales. It was shown that the Mount Wilson classification encodes the notion of activity over all length scales in the active region, and there are significant differences in the gradient distribution between flaring and non-flaring active regions over all length scales. However, timescales of the sequence of predictors have not been analyzed. Furthermore, the sequence of predictors representing the evolution of emerging flux regions was used to forecast solar flares (Yu et al. 2009); however, the influence of the changing rate of emerging fluxes on flare prediction has not been considered. Here, maximum overlap discrete wavelet transform is used to decouple the sequence of predictors in different frequency bands in which a different changing rate of the sequence of predictors is reflected. The statistic characteristics of multiresolution predictors are analyzed. Based on these predictors, a short-term solar flare prediction model is built by the C4.5 decision tree algorithm. The performances of the proposed model are analyzed.

The rest of this paper is organized as follows. The data are introduced in Section 2. The information gain ratio of multiresolution predictors is calculated in Section 3. The performances of the proposed model are analyzed in Section 4. Finally, the conclusions are given in Section 5.

## 2. DATA

The general description of the data can be found in Cui et al. (2006) and Yu et al. (2009). Flare data are downloaded from http://www.ngdc.noaa.gov/stp/SOLAR/ftpsolarflares.html#xray. Within a certain interval, more than one flare may occur. The importance of these flares is summed up with weights. The total importance of flares is defined as follows:

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

Equation (1) considers the influence of all the flares within the forward-looking period. The influence of different forward-looking periods has been discussed in Cui et al. (2006). Here, this period is 48 hr. A forecasting model usually pays attention to the production of flares with significance above a threshold. The threshold of $I_{\rm tot}$ is supposed to be 10 in the present work. It means that the definition of "flaring" versus "non-flaring" is a total importance above M1.0.

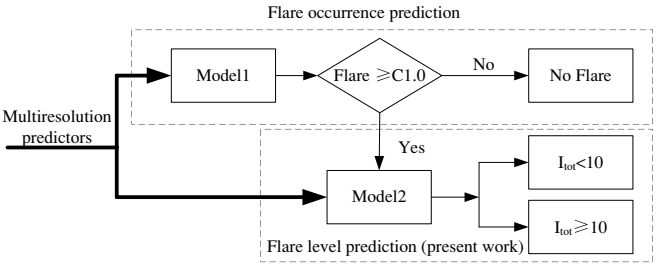


**Figure 1.** Framework of a short-term solar prediction system.

The maximum horizontal gradient ($|\nabla_h B_z|_m$), the length of neural line (L), and the number of singular points ($\eta$) extracted from active regions in *SOHO*/MDI full disk longitudinal magnetograms are used as predictors of solar flares. Active region location data associated with the solar flare events are obtained from Solar Geophysical Data solar event reports (http://www.solarmonitor.org/index.php). Two criteria are used to select the active regions.

1. At least one X-ray flare whose magnitude ⩾C1.0 is produced. This criterion means that we only focus on the active regions producing at least one C1.0 flare. However, it ensures that the present work cannot be used in real-time application. If this work needs to be applied in real-time prediction, the other prediction model called Model 1 in Figure 1, which can be built by the data-driven method or the experience of experts, should be built first. As shown in Figure 1, when Model 1 predicts that flares above C1.0 will be produced, the present work is used to forecast whether $I_{\rm tot}$ of these flares is larger than M1.0. This framework is similar to that proposed by Qahwaji & Colak (2007). The performance of Qahwaji & Colak (2007) shows that flare occurrence prediction is easier than flare level prediction, so the present work focuses on whether the $I_{\rm tot}$ is larger than the settled threshold.
2. The location of active regions is within 30° of the solar disk center where projection effects can be negligible. However, all the corresponding flares over the disk transit are considered.

Data from 1996 April 15 to 2008 April 2 are used to analyze the relationships between the magnetic field measures and the flares. According to the above two criteria, 29,772 magnetograms containing 1010 active regions are extracted, and these active regions appeared in 29,772 maps 55,582 times. Meanwhile, 5241 flares are included; among those are 4517 C level flares, 649 M level flares, and 75 X level flares.

Yu et al. (2009) introduced the sequence of predictors by the sliding window method:

$$\mathbf{x}(t - W\Delta t) \quad \cdots \quad \mathbf{x}(t - \Delta t) \quad \mathbf{x}(t) \quad I_{\rm tot}(t + F)\ , \qquad (2)$$

where $\mathbf{x}(t)$ is the vector of predictors at time $t$, $\mathbf{x} = \{|\nabla_h B_Z|_m\ , L, \eta\}$, $\Delta t$ is 96 minutes, which is the interval between the successive magnetograms, $W$ is the length of the sequence, $F$ is the forecasting time, and $I_{\rm tot}(t + F)$ is the total importance of flares within the interval $F$. When $I_{\rm tot} = 10$ and $F = 48$(hr), the appropriate length of the sequence is 45 ($W = 45$). At the beginning of the observation, the first value of samples is repeated $W$ times to construct the sequence (Yu et al. 2009).

All the predictors are pre-processed by the sigmoid function to incorporate the prior information into the machine learning

**Table 1**
Values of Parameters in Boltzmann Functions

| Threshold | Forward-looking Period | Predictor | $A_1$ | $A_2$ | $X_0$ | W |
|---|---|---|---|---|---|---|
| $I_{\rm tot} = 10$ | 48(hr) | $\lvert\nabla_h B_z\rvert_m$ | 0.164 | 0.738 | 0.360 | 0.066 |
| | | L | 0.062 | 0.848 | 763.08 | 382.97 |
| | | $\eta$ | −0.196 | 0.730 | 9.343 | 22.663 |

algorithm. The sigmoid relationship between predictors and the flare productivity found by Cui et al. (2006) is used to reduce the complexity of the prediction model and improve the generalization of this model. The sigmoid function in the Boltzmann style is as follows:

$$Y = A_2 + \frac{A_1 - A_2}{1 + \exp[(X - X_0)/W]}, \quad (3)$$

where $Y$ is the flare productivity defined by the ratio of the number of flaring samples to the number of total samples, and $X$ is the value of the predictor. $A_1$, $A_2$, $X_0$, and $W$ are estimated from the curve-fitting process. In this process, parameters $A_1$, $A_2$, $X_0$, and $W$ are optimized to minimize the sum of the squares of the deviations between the observed data and the expected data (Marko 2003). The values of these parameters are given in Table 1. The physical meaning of different shapes of sigmoid functions is discussed in Wang et al. (2009).

## 3. MULTIRESOLUTION PREDICTORS

### 3.1. Construction of Multiresolution Predictors

The sequence of predictors reflects the evolution of emerging flux regions. However, the changing rate of emerging fluxes has not been considered. Predictors in different frequency bands can provide different information about the evolutionary velocity of emerging fluxes. The information in different frequency bands is decoupled by wavelet transform for the sequence of predictors. Continuous wavelet transform is an exploratory multiresolution analysis tool for time series. However, it is redundant since the two-dimensional continuous wavelet transform depends on just a one-dimensional signal (Percival & Walden 2000). Discrete wavelet transform preserves the key features by sub-sampling continuous wavelet transform for both time and scale. However, the discrete wavelet transform of level $J_0$ restricts the sample size to an integer multiple of $2^{J_0}$, and its results are different for the time series with different starting points. Here, the length of the sequence of predictors is 45, and the sequence of predictors shifts with the sampling interval in each active region. So maximal overlap discrete wavelet transform which does not sub-sample for time is used to decompose the sequences of predictors into different frequency bands. It is the balance of continuous wavelet transform and discrete wavelet transform. After the sequence is decomposed into four frequency bands, four sequential features shown in Table 2 are extracted in each frequency band to construct the multiresolution predictors.

### 3.2. Information Gain Ratio of Multiresolution Predictors

Entropy is used to measure the uncertainty of a system. For the solar flare prediction system, $F$ ($F = \{f_1, f_2, \ldots, f_k\}$) and $O$ ($O = \{o_1, o_2, \ldots, o_l\}$) are used to stand for the flare level and the magnetic field observation of the active region, respectively.

**Table 2**
Definitions of Sequential Features for $\mathbf{y} = \{y_1, y_2, \ldots, y_W\}$

| Feature | Definition |
|---|---|
| Maximum | $\mathrm{Max}(\mathbf{y}) = \{y_i \mid y_i \geqslant y_j, \forall y_j \in \mathbf{y}\}$ |
| Mean | $\mathrm{Mean}(\mathbf{y}) = \frac{1}{W}\sum_{i=1}^{W} y_i$ |
| Standard deviation | $\mathrm{Std}(\mathbf{y}) = \sqrt{\frac{1}{W}\sum_{i=1}^{W}(y_i - \mathrm{mean}(\mathbf{y}))^2}$ |
| Root mean square | $\mathrm{RMS}(\mathbf{y}) = \sqrt{\frac{1}{W}\sum_{i=1}^{W} y_i^2}$ |

The entropy of $F$ is

$$H(F) = -\sum_{i=1}^{k} P(F = f_i)\log_2(P(F = f_i)). \quad (4)$$

The conditional entropy of $F$ given $O$ is

$$H(F|O) = -\sum_{j=1}^{l} P(O = o_j)H(F|O = o_j). \quad (5)$$

In information theory, the decrease of uncertainty is called information which is quantitatively measured by information gain (IG).

$$\mathrm{IG}(F, O) = H(F) - H(F|O). \quad (6)$$

$\mathrm{IG}(F, O)$ means the decrease of uncertainty of $F$, when observation $O$ is given. However, information gain biases the predictor with a large number of distinct values, so the information gain ratio (GR) is defined to solve this drawback.

$$\mathrm{GR}(F, O) = \frac{\mathrm{IG}(F, O)}{H(O)}. \quad (7)$$

For example, $\mathrm{GR}(F, O_1) = 0.06$ means that after eliminating the influence of the number of distinct values in $O_1$, the uncertainty of $F$ decreases 0.06 because of the observation of $O_1$. If $\mathrm{GR}(F, O_2) = 0.02$, this means that the decrease of uncertainty given $O_2$ is less than the decrease of uncertainty given $O_1$. Namely, the information of $O_2$ is less than the information of $O_1$ for the flare prediction.

In order to estimate the information gain ratio between multiresolution predictors and solar flares, these predictors are discretized by the entropy based discretization method (Fayyad & Irani 1993), and then the information gain ratio of all the multiresolution predictors is shown in Figure 2. It is found that the predictors within the lowest frequency band and the highest frequency band contain the most information for flare prediction. The multiresolution predictors in the lowest frequency band represent the evolutionary tendency of emerging flux regions, and the multiresolution predictors in the highest frequency band reflect the furthest variation of the emerging fluxes. It is concluded that not only the evolutionary

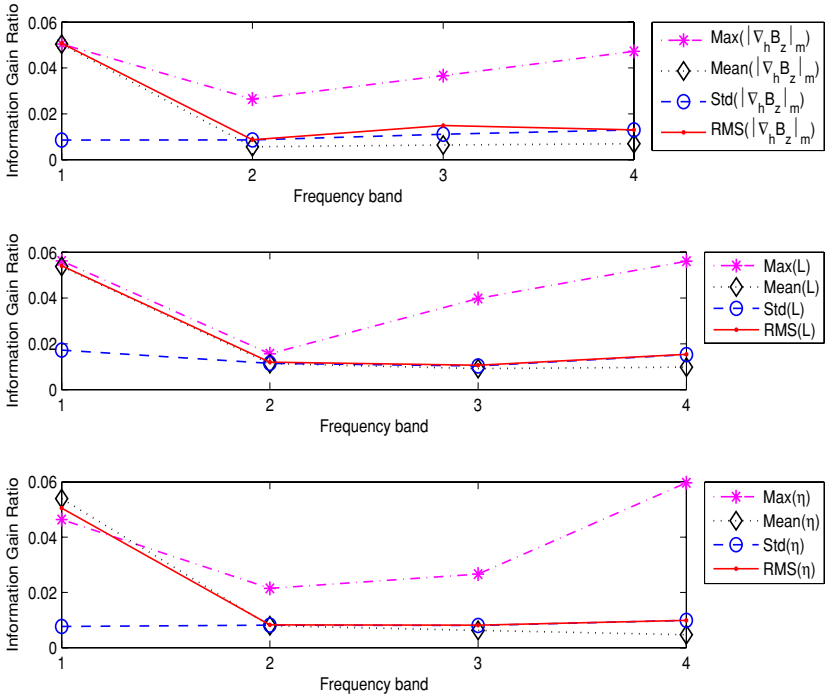


**Figure 2.** Information gain ratio of multiresolution predictors within four frequency bands. 1, 2, 3, and 4 labeled in the X-axes stand for the frequency bands 1–4 which are (0–1/8) $f_{max}$, (1/8–1/4) $f_{max}$, (1/4–1/2) $f_{max}$, and (1/2–1) $f_{max}$, respectively. $f_{max}$ is the highest frequency of the sequence.
(A color version of this figure is available in the online journal.)

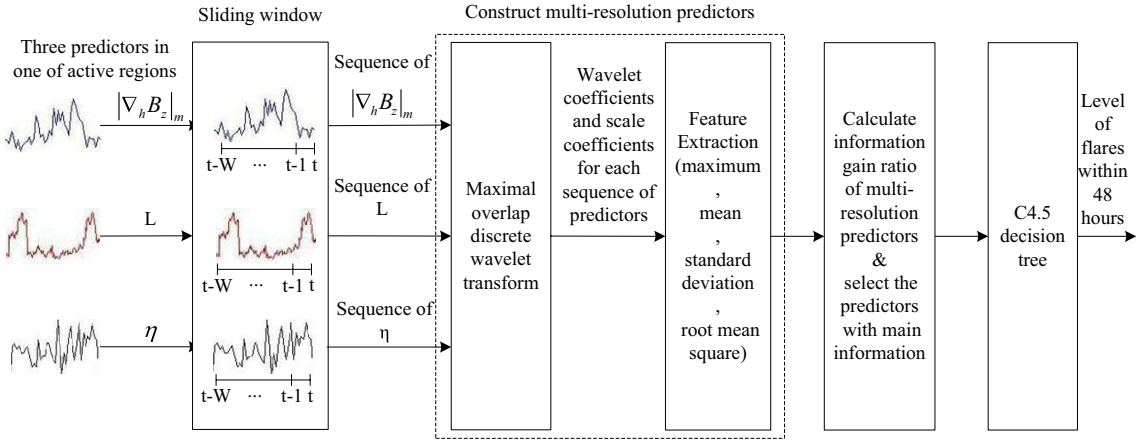


**Figure 3.** General schematic view of flare level prediction using multiresolution predictors.
(A color version of this figure is available in the online journal.)

tendency of emerging flux regions but also the furthest changing rate of emerging fluxes are important for flare prediction, so the multiresolution predictors within the lowest and highest frequency band are used to generate a short-term solar flare prediction model by the C4.5 decision tree algorithm (Quinlan 1993). The application of the C4.5 decision tree in solar flare prediction can be found in Yu et al. (2009).

A general schematic view of flare level prediction based on multiresolution predictors is shown in Figure 3. The sequences of predictors are generated by the sliding window method (Yu et al. 2009), and then multiresolution predictors are constructed by the wavelet transform and sequential feature extraction method. The information amount of the proposed predictors is measured by the information gain ratio. The lowest and highest frequency bands contain the most information, so these predictors are selected as the input of the C4.5 decision tree. Finally, the flare level will be forecast by the trained model.

**Table 3**
Different Outcomes of Two-class Prediction

| Class of Samples | Predicted Positive Class | Predicted Negative Class |
|---|---|---|
| Actual positive class | True positive | False negative |
| Actual negative class | False positive | True negative |

## 4. EXPERIMENTAL RESULTS AND ANALYSES

### *4.1. Performance Evaluation*

The results of the proposed prediction model are grouped into "flaring" or "non-flaring." The flaring samples are considered either positive classor negative class. In this case, the prediction model has four different possible outcomes as shown in Table 3.

The samples correctly classified as positive are defined as true positive (TP), while the samples correctly classified as negative

**Table 4**
The Confusion Matrices Generated by C4.5 Decision Tree Based on Different Predictors

| No. of Samples | Raw Sequence | | DWT_Haar | | DWT_DB2 | | | |
|---|---|---|---|---|---|---|---|---|
| | PP | PN | PP | PN | PP | PN | | |
| AP | 836 ± 12 | 144 ± 12 | 856 ± 9 | 124 ± 9 | 851 ± 10 | 130 ± 10 | | |
| AN | 733 ± 27 | 3845 ± 27 | 735 ± 48 | 3843 ± 48 | 711 ± 48 | 3868 ± 48 | | |
| | MODWT_Haar | | MODWT_DB2 | | MODWT_Haar_Red | | MODWT_DB2_Red | |
| | PP | PN | PP | PN | PP | PN | PP | PN |
| AP | 925 ± 6 | 55 ± 6 | 924 ± 9 | 57 ± 9 | 930 ± 8 | 50 ± 8 | 927 ± 7 | 53 ± 7 |
| AN | 383 ± 16 | 4195 ± 16 | 396 ± 30 | 4182 ± 30 | 387 ± 16 | 4191 ± 16 | 371 ± 20 | 4207 ± 20 |

**Notes.** Confusion matrices for prediction model based on multiresolution predictors constructed by discrete wavelet transform with Haar wavelet basis (DWT_Haar), discrete wavelet transform with DB2 wavelet basis (DWT_DB2), maximum overlap discrete wavelet transform with Haar wavelet basis (MODWT_Haar), maximum overlap discrete wavelet transform with DB2 wavelet basis (MODWT_DB2), maximum overlap discrete wavelet transform with Haar wavelet basis in the lowest and highest frequency bands (MODWT_Haar_Red), and maximum overlap discrete wavelet transform with DB2 wavelet basis in the lowest and highest frequency bands (MODWT_DB2_Red). PP stands for predicted positive class, PN stands for predicted negative class, AP stands for actual positive class, and AN stands for actual negative class.

**Table 5**
Performance Comparisons of Prediction Models Based on Different Predictors

| Performance Evaluation | Raw Sequence | DWT_Haar | DWT_DB2 | |
|---|---|---|---|---|
| TP rate (%) | 85.3 ± 1.3 | 87.4 ± 0.9 | 86.8 ± 1.0 | |
| TN rate (%) | 84.0 ± 0.6 | 83.9 ± 1.1 | 84.5 ± 1.1 | |
| HSS | 0.56 ± 0.02 | 0.57 ± 0.02 | 0.58 ± 0.02 | |
| | MODWT_Haar | MODWT_DB2 | MODWT_Haar_Red | MODWT_DB2_Red |
| TP rate (%) | 94.4 ± 0.6 | 94.2 ± 0.9 | 94.9 ± 0.8 | 94.6 ± 0.7 |
| TN rate (%) | 91.6 ± 0.3 | 91.4 ± 0.6 | 91.5 ± 0.3 | 91.9 ± 0.4 |
| HSS | 0.76 ± 0.01 | 0.75 ± 0.02 | 0.76 ± 0.01 | 0.77 ± 0.01 |

**Notes.** The definition of prediction models is the same as the definition in Table 4.

are defined as true negative (TN). On the other hand, the samples wrongly predicted as positive are defined as false positive (FP) and the samples wrongly predicted as negative are defined as false negative (FN). Prediction performance is measured using the TP and TN rates.

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:

$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

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:

$$\text{TN rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{9}$$

The TP and TN rates are used to measure the performance of flaring and non-flaring predictions, respectively. Furthermore, the Heidke skill score (HSS) is used to generally evaluate the performance of the proposed method (Jolliffe & Stephenson 2003).

$$\text{HSS} = \frac{\text{PC} - \text{E}}{1 - \text{E}}. \tag{10}$$

where $\text{N} = \text{TP+TN+FP+FN}$, $\text{PC} = \frac{\text{TP+TN}}{\text{N}}$, and $\text{E} = \frac{(\text{TP+FN})(\text{TP+FP})}{\text{N}^2} + \frac{(\text{TN+FP})(\text{TN+FN})}{\text{N}^2}$. E is PC for random forecast, so this version of HSS shows the increase in predictive power over that of random chance.

### *4.2. Performance of Flare Prediction Model*

The data set contains 9801 flaring samples and 45,781 non-flaring samples. This is called the class imbalance problem in the data mining community. The model derived from the unbalanced data set will be biased toward the non-flaring samples. In the present work, the training set is undersampled to balance its class distribution (Japkowicz & Stephen 2002). The data set is divided into tenfolds and then ninefolds are used for training, and the remaining fold for testing. The performances of the prediction model are given as the mean of all tests, and the uncertainty of the results is estimated by its standard deviation. The prediction models are generated by the C4.5 decision tree algorithm implemented in Waikato Environment for Knowledge Analysis (Witten & Frank 2005).

The confusion matrices generated by the C4.5 decision tree based on different predictors are shown in Table 4. The other verification measures can be calculated from these matrices.

The performances of models based on predictors with different construction methods are shown in Table 5. By comparing the performances of these models, the following conclusions can be obtained.

First, the performance of the model based on the multiresolution predictors is better than that based on the raw sequence. There are two main reasons why the performance of the proposed prediction model is improved. On the one hand, the information for the changing rate of emerging fluxes provided by the multiresolution predictors is valuable for forecasting solar flares. On the other hand, the evolutionary information of emerging flux regions in different frequency bands is decoupled by wavelet transform, so that the generalization ability of the prediction model is improved.

Second, the performances of the model based on multiresolution predictors constructed by maximum overlap discrete wavelet transform are better than the performance of the model based on multiresolution predictors constructed by discrete wavelet transform. Because the decomposed coefficients of discrete wavelet transform are sensitive to the starting point of sequences of predictors, this will influence the stability of multiresolution predictors. It is concluded that maximum overlap discrete wavelet transform is more suitable than discrete wavelet transform for constructing multiresolution predictors.

Third, the performances of the model based on Haar wavelet and DB2 wavelet are almost the same, although the wavelet-based algorithms may introduce spurious features in the transform. It is found that the multiresolution predictors are not sensitive to the selection of wavelet basis functions.

Finally, the performance of the model based on the multiresolution predictors in the lowest and highest predictors is comparable with the performance of the model based on all the multiresolution predictors. This means that the multiresolution predictors extracted form the lowest and highest frequency bands maintain the most information. It is consistent with the analysis of the information gain ratio.

## 5. CONCLUSIONS

From the point of view of information provided, the multiresolution predictors in four frequency bands reflect the trend and the changing rate of emerging flux regions. This information is important for short-term solar flare prediction. From the point of view of generalization ability, the variation of the emerging fluxes is decoupled with the wavelet transform, and the generalization ability of the short-term solar flare prediction model using the multiresolution predictors is improved. From the point of view of construction of multiresolution predictors, maximum overlap discrete wavelet transform is more suitable than discrete wavelet transform for generating the multiresolution predictors, because it can treat the sequence with any number of sampling points, and it is shift invariant for the sequence of predictors.

The information gain ratio of all multiresolution predictors is calculated. It quantitatively measures the information amount provided by multiresolution predictors. By comparing the information gain ratio of predictors within different frequency bands, it is found that predictors in the lowest and highest frequency bands contain more information to predict flares. This is further approved by comparing the performance of the model based on the multiresolution predictors extracted from the lowest and highest frequency bands and that based on all the multiresolution predictors.

The performance of models based on predictors constructed by different wavelet basis functions is comparable; the result shows that the prediction model is not sensitive to the selection of wavelet basis functions. This is an attractive characteristic for the constructed multiresolution predictors. The performance of the present method increased 9% for the TP rate and 7% for the TN rate over the prediction model based on the raw sequence of predictors. This indicates that the short-term solar flare prediction model using the multiresolution predictors will play an important role in forecasting services in the future.

The present work can be used to predict large flares once a small flare has occurred. This makes the proposed algorithm less useful for real-time application. In order to solve this problem, an additional model should be built to predict whether the active region will produce a small flare. Once this model is built in the future, this combined system can be applied for real-time prediction.

This work is supported by the National Natural Science Foundation of China under grant nos. 10978011, 10673017, and 10733020, and the National Basic Research Program of China (973 Program) under grant no. 2006CB806307. This paper has been greatly improved according to the suggestions of the anonymous reviewer. In addition, we thank the *SOHO*/MDI consortium for the data. *SOHO* is a project of international cooperation between ESA and NASA.

## REFERENCES

Abramenko, V. I. 2005, ApJ, 629, 1141
Barnes, G., Leka, K. D., Schumer, E. A., & Della-Rose, D. J. 2007, Space Weather, 5, S09002
Bornmann, P. L., & Shaw, D. 1994, Sol. Phys., 150, 127
Bradshaw, G., Fozzard, R., & Ceci, L. 1989, Adv. Neural Inf. Process., 1, 248
Cui, Y. M., Li, R., Wang, H. N., & He, H. 2007, Sol. Phys., 242, 1
Cui, Y. M., Li, R., Zhang, L. Y., He, Y. L., & Wang, H. N. 2006, Sol. Phys., 237, 45
Delouille, V., De Patoul, J., Hochedez, J. F., Jacques, L., & Antoine, J. P. 2005, Sol. Phys., 228, 301
Falconer, D. A. 1997, Sol. Phys., 176, 123
Falconer, D. A., Moore, R. L., & Gary, G. A. 2006, ApJ, 644, 1258
Fayyad, U. M., & Irani, K. B. 1993, in Proc. of the 13th International Joint Conf. on Artificial Intelligence, Chambery, France, ed. R. Bajcsy (San Francisco, CA: Morgan–Kaufmann), 1022
Gallagher, P. T., Moon, Y. J., & Wang, H. 2002, Sol. Phys., 209, 171
Georgoulis, M. K., & Rust, D. M. 2007, ApJ, 661, 109
He, H., Wang, H. N., Du, Z. L., Li, R., Cui, Y. M., Zhang, L. Y., & He, Y. L. 2008, Adv. Space Res., 42, 1450
Hewett, R. J., Gallagher, P. T., McAteer, R. T. J., Young, C. A., Ireland, J., Conlon, P. A., & Maguire, K. 2008, Sol. Phys., 248, 311
Ireland, J., Young, C. A., McAteer, R. T. J., Whelan, C., Hewett, R. J., & Gallagher, P. T. 2008, Sol. Phys., 252, 121
Japkowicz, N., & Stephen, S. 2002, Intell. Data Anal., 6, 429
Jing, J., Song, H., Abramenko, V., Tan, C., & Wang, H. 2006, ApJ, 644, 1273
Jolliffe, I. T., & Stephenson, D. B. 2003, Forecast Verification: A Practitioner's Guide in Atmospheric Science (New York: Wiley)
Leka, K. D., & Barnes, G. 2003, ApJ, 595, 1277
Leka, K. D., & Barnes, G. 2007, ApJ, 656, 1173
Li, R., Wang, H. N., He, H., Cui, Y. M., & Du, Z. L. 2007, Chin. J. Astron. Astrophys., 7, 441
Marko, L. 2003, Ind. Phys., 9, 24
McAteer, R. T. J., Gallagher, P. T., & Ireland, J. 2005, ApJ, 631, 628
McIntosh, P. S. 1990, Sol. Phys., 125, 251
Miller, R. W. 1989, in Knowledge-Based Systems in Astronomy, ed. A. Heck & F. Murtagh (Lecture Notes in Physics, Vol. 329; Berlin: Springer), 1616
Moore, R. L., Sterling, A. C., Hudson, H. S., & Lemen, J. R. 2001, ApJ, 552, 833
Percival, D. B., & Walden, A. T. 2000, Wavelet Methods for Time Series Analysis (Cambridge: Cambridge Univ. Press)
Portier-Fozzani, F., Vandame, B., Bijaoui, A., Maucherat, A. J., & EIT Team 2001, Sol. Phys., 201, 271
Qahwaji, R., & Colak, T. 2007, Sol. Phys., 241, 195
Quinlan, J. R. 1993, C4.5: Programs for Machine Learning (San Mateo, CA: Morgan Kaufmann Publishers)
Sammis, I., Tang, F., & Zirin, H. 2000, ApJ, 540, 583
Schrijver, C. J. 2007, ApJ, 655, 117
Wang, H. N., Cui, Y. M., & Han, H. 2009, Res. Astron. Astrophys., 9, 687
Wang, H. N., Cui, Y. M., Li, R., Zhang, L. Y., & Han, H. 2008, Adv. Space Res., 42, 1464
Wheatland, M. S. 2004, ApJ, 609, 1134
Witten, I. H., & Frank, E. 2005, Data Mining: Practical Machine Learning Tools and Techniques (San Mateo, CA: Morgan Kaufmann Publishers)
Yu, D. R., Huang, X., Wang, H. N., & Cui, Y. M. 2009, Sol. Phys., 255, 91