The Astrophysical Journal, 710:869–877, 2010 February 10 doi:10.1088/0004-637X/710/1/869

SHORT-TERM SOLAR FLARE LEVEL PREDICTION USING A BAYESIAN NETWORK APPROACH

Daren Yu[1], Xin Huang[1], Huaning Wang[2], Yanmei Cui[3], Qinghua Hu[1], and Rui Zhou[1]
[1] Harbin Institute of Technology, No. 92 West Da Zhi Street, Harbin, Heilongjiang Province 150001, China; huangxinhit@yahoo.com.cn
[2] National Astronomical Observatories, 20A Datun Road, Chaoyang District, Beijing 100012, China
[3] Center for Space Science and Applied Research, No. 1 Nanertiao, Zhongguancun, Haidian District, Beijing 100080, China
Received 2009 March 23; accepted 2010 January 6; published 2010 January 22

ABSTRACT

A Bayesian network approach for short-term solar flare level prediction has been proposed based on three sequences of photospheric magnetic field parameters extracted from *Solar and Heliospheric Observatory*/Michelson Doppler Imager longitudinal magnetograms. The magnetic measures, the maximum horizontal gradient, the length of neutral line, and the number of singular points do not have determinate relationships with solar flares, so the solar flare level prediction is considered as an uncertainty reasoning process modeled by the Bayesian network. The qualitative network structure which describes conditional independent relationships among magnetic field parameters and the quantitative conditional probability tables which determine the probabilistic values for each variable are learned from the data set. Seven sequential features—the maximum, the mean, the root mean square, the standard deviation, the shape factor, the crest factor, and the pulse factor—are extracted to reduce the dimensions of the raw sequences. Two Bayesian network models are built using raw sequential data (BN_R) and feature extracted data (BN_F), respectively. The explanations of these models are consistent with physical analyses of experts. The performances of the BN_R and the BN_F appear comparable with other methods. More importantly, the comprehensibility of the Bayesian network models is better than other methods.

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

Online-only material: color figures

1. INTRODUCTION

Large solar flares are associated with various effects of space weather (Koskinen et al. 1999), so it is important to predict their eruptions.

It is accepted that photospheric morphology of active regions is related to a solar flare. McIntosh (1990) introduced the definitions of the McIntosh classification. Based on these classifications of sunspots, an expert system (Theo) was built to predict X-ray solar flares, and Bradshaw et al. (1989) constructed a three-layer back-propagation neural network named TheoNet to forecast flares. Bornmann & Shaw (1994) pointed out that the McIntosh parameters act as proxies for the magnetic properties of the active region and discussed the relationships among them. It was concluded that the first parameter provides a measure of the total magnetic flux within the active region, the second parameter provides a measure of the magnetic flux of the largest spot, and the third parameter serves as a measure of the total area of sunspots. Gallagher et al. (2002) estimated the daily flare rate based on the McIntosh classification of the active region, and then the probability of occurrence for one or more M class flares was modified according to the assumption of a constant Poisson process for the waiting-time distribution of X-ray flares. Li et al. (2007) combined support vector machine and *K*-nearest neighbors to construct a solar flare forecasting model. Qahwaji & Colak (2007) built a hybrid system which combines a support vector machine and a cascade-correlation neural network for automatic short-term solar flare prediction. Colak & Qahwaji (2008) presented a hybrid system for automatic detection and McIntosh-based classification of sunspot groups. Based on the work of Qahwaji & Colak (2007) and Colak & Qahwaji (2008), Colak & Qahwaji (2009) presented an automated hybrid computer platform (ASAP) for short-term prediction of significant solar flares using *Solar and Heliospheric Observatory* (*SOHO*)/Michelson Doppler Imager (MDI) images.

Lots of efforts have been made to directly find the relationships between magnetic field properties and flares (Leka & Barnes 2003a, 2003b; McAteer et al. 2005; Jing et al. 2006; Cui et al. 2006, 2007; Schrijver 2007; Georgoulis & Rust 2007). Based on the parameters proposed by Leka & Barnes (2003a), Leka & Barnes (2007) applied the Fisher's linear discriminant analysis to identify whether a flare will happen. Furthermore, this method was extended to probabilistic prediction by Barnes et al. (2007). Based on the parameters proposed by Cui et al. (2006), Wang et al. (2008) trained a neural network for solar flare prediction. Yu et al. (2009, 2010) analyzed the influence of sequences of magnetic filed parameters on the flare level, and then the solar flare prediction model was built under the sequential supervised learning framework.

Assuming that flares obey Poisson distribution in time and power-law distribution in size, Wheatland (2001) presented a method of solar flare prediction using the observed flare statistics. According to these phenomenological rules and the flaring records, Wheatland (2004) proposed a Bayesian approach to flare prediction. The prior probability of the prediction was given, and then the flaring records together with phenomenological rules of flare statistics were used to refine the initial prediction. It was stated that this method is simple, objective, and makes few ad hoc assumptions (Wheatland 2005). However, this method ignores the valuable information contained in the magnetic field of active regions. Here, the Bayesian network encoded the conditional independent relationships among the magnetic field properties of the active region to predict the flare level. In the hybrid solar flare prediction system proposed by Qahwaji & Colak (2007), the whole prediction of solar flares contained two main parts: flare occurrence prediction and flare

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
		η	−0.196	0.730	9.343	22.663

level prediction. Flare occurrence prediction is used to forecast whether an active region will produce a flare, and if so, flare level prediction is used to forecast whether the flare is going to be above a certain threshold. The present work focuses on the flare level prediction. In the process of flare level prediction, seven features are extracted from the sequences of magnetic field properties. Furthermore, the temporal variations of these features are related with the eruption of flares. The physical explanations of the Bayesian network models are given and the performances of these models are compared with other methods.

The rest of this paper is organized as follows. The data are introduced in Section 2. The sequential features and their varieties with the eruption of flares are analyzed in Section 3. The Bayesian network and the generated prediction model are described in Section 4. The experimental results and comparisons are presented in Section 5. Finally, the conclusions and discussions are given in Section 6.

2. DATA

2.1. Original Data

The maximum horizontal gradient ($\lvert\nabla_h B_z\rvert_m$), the length of neural line (L), and the number of singular points (η) of the photospheric magnetic field in active regions are calculated as preflare signatures called predictors in the flare prediction model (Cui et al. 2006). They are extracted from *SOHO*/MDI full disk longitudinal magnetograms with a pixel size of 2″ and noise level of 20 G from 1996 April 15 to 2004 January 10. The interval between the successive magnetograms is 96 minutes.

Active region location data associated with the solar flare events are obtained from Solar Geophysical Data (SGD) solar event reports (http://www.solarmonitor.org/index.php). Active regions are selected using the following two criteria:

1. At least one X-ray flare whose magnitude $\geqslant$C1.0 is produced in these active regions.
2. The location of active regions is within 30° of the solar disk center.

Generally the large flares are paid more attention. Therefore, criterion one is proposed to focus on the active regions above the certain threshold. Criterion two is used to reduce the impact of projection effects. The active regions are extracted by hand. A rectangular patch is used to select the active region. When two active regions are in the same patch, they are considered as one active region.

Flare data are downloaded from http://www.ngdc.noaa.gov/stp/SOLAR/ftpsolarflares.html#xray. The importance of a solar flare is conventionally described by its index, for example, C, M, or X. Within the forecasting period, more than one flare may happen. The importance of these flares is summed up with weights. The total importance of flares is computed 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. For example, if an active region produces C1.2, C2.3, M4.1, and X1.2 flares within 48 hr, we have $I_{\rm tot} = (1.2 + 2.3) + 10 \times 4.1 + 100 \times 1.2 = 164.5$ (Wang et al. 2008). A forecasting model usually pays attention to the production of flares with significance above a threshold. Here, the threshold of $I_{\rm tot}$ was chosen to be 10. Thus, the non-flaring sample is defined to have a total importance less than 10.

The predictors are pre-processed to incorporate prior information into machine learning algorithms. The prior information is the known relationships between the inputs and the outputs of a system. For the flare prediction system, it is the known relationships between the predictors and the flare level. Cui et al. (2006) point out that the relationships between the predictors and the flare productivity obey the sigmoid function in Boltzmann style shown in Equation (2). So the predictors are mapped by this function, and then the Bayesian network is used to obtain the other relationships between the predictors and the flare level from the data set.

$$Y = A_2 + \frac{A_1 - A_2}{1 + \exp[(X - X_0)/W]} \qquad (2)$$

where Y is the flare productivity defined by the ratio of the number of flare-productive 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 with 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.

2.2. Sequences of Predictors

Yu et al. (2009) introduced the evolutionary information of predictors by the sliding window method and analyzed its influence on the flare level. The sequence of the predictors is represented as

$$\mathbf{x}(t)\ \mathbf{x}(t - \Delta t) \cdots \mathbf{x}(t - W\Delta t)\, I_{\rm tot}(t + F), \qquad (3)$$

where $\mathbf{x}(t)$ is the vector of predictors at time t. $\mathbf{x} = \{\lvert\nabla_h B_Z\rvert_m, L, \eta\}$. Δt is the interval between two observations. W is the length of the sequences. F is the forecasting time. $I_{\rm tot}(t + F)$ is the total importance of flares within the interval F.

The evolutionary information is contained in the sequence of predictors. Yu et al. (2009) pointed out that the appropriate value of W is 45 for each predictor. There are not enough data for the sliding window at the beginning of the observation of an active region. So the first observational value is repeated W times to provide the initial values. For the missing magnetogram, the previous magnetogram is repeated instead of the missing one.

Table 2
Definitions of Sequential Features

Feature	Definition	Characteristics
Maximum	$\mathrm{Max}(\mathbf{y}) = \{y_i \mid y_i \geqslant y_j, \forall y_j \in \mathbf{y}\}$	Sensitive
Crest factor	$\mathrm{Crest}(\mathbf{y}) = \frac{\mathrm{Max}(\mathbf{y})}{\mathrm{rms}(\mathbf{y})}$	Sensitive
Pulse factor	$\mathrm{Pulse}(\mathbf{y}) = \frac{\mathrm{Max}(\mathbf{y})}{\mathrm{Mean}(\mathbf{y})}$	Sensitive
Mean	$\mathrm{Mean}(\mathbf{y}) = \frac{1}{W}\sum_{i=1}^{W} y_i$	Robust
Root mean square	$\mathrm{rms}(\mathbf{y}) = \sqrt{\frac{1}{W}\sum_{i=1}^{W} y_i^2}$	Robust
Standard deviation	$\mathrm{Std}(\mathbf{y}) = \sqrt{\frac{1}{W}\sum_{i=1}^{W}(y_i - \mathrm{mean}(\mathbf{y}))^2}$	Robust
Shape factor	$\mathrm{Shape}(\mathbf{y}) = \frac{\mathrm{rms}(\mathbf{y})}{\mathrm{Mean}(\mathbf{y})}$	Robust

3. FEATURE EXTRACTION FOR THE SEQUENCES OF PREDICTORS

For each type of predictor x (x stands for $|\nabla_h B_Z|_m$, L or η), its sequence is noted as

$$\mathbf{y} = \{x(t), x(t - \Delta t), \ldots, x(t - W\Delta t)\}. \quad (4)$$

Seven features—the maximum, the mean, the root mean square, the standard deviation, the shape factor, the crest factor, and the pulse factor—are extracted from the sequences of $|\nabla_h B_Z|_m$, L, and η, respectively. Their definitions are listed in Table 2. Because the maximum of a sequence is easily influenced by a single point with the large value, the maximum and its related features are sensitive to changes in a single measurement. The other predictors are robust for the variation of a single point.

The variations of the features of maximum horizontal gradient are shown in Figures 1 and 2. Three active regions (AR 7978, AR 9373, and AR 9494) are selected. At the beginning of these active regions, the observational samples are non-flaring and at the end of these active regions, the observational samples are flaring, the boundary is plotted as the first vertical solid line. The forecasting time is 48 hr indicated by the second vertical solid line. During the forecasting time, several flares happened. The eruption of a C or M level flare is indicated by the dash-dotted line or dashed line, respectively. The corresponding flare level is labeled beside these vertical lines. As shown in Figure 1, the maximum, the pulse factor, and the crest factor undergo abrupt changes, so they are sensitive to changes in a single measurement. In Figure 2, the mean, the root mean square, the standard deviation, and the shape factor vary smoothly. They are robust to changes in a single measurement.

4. BAYESIAN NETWORK APPROACH

Bayesian reasoning provides a probabilistic approach to inference (Mitchell 1997). A prior probability of a hypothesis provides its initial knowledge, and then the observed data are used to refine the reasoning process. Bayes' formula is as follows:

$$P(y|\mathbf{X}) = \frac{P(\mathbf{X}|y)P(y)}{P(\mathbf{X})}, \quad (5)$$

where y is the decision and $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is the set of predictors.

Using the chain rule of probability, the joint distribution $P(\mathbf{X}|y)$ is factored as

$$P(\mathbf{X}|y) = \prod_{i=1}^{n} P(x_i|y, x_1, x_2, \ldots, x_{i-1}). \quad (6)$$

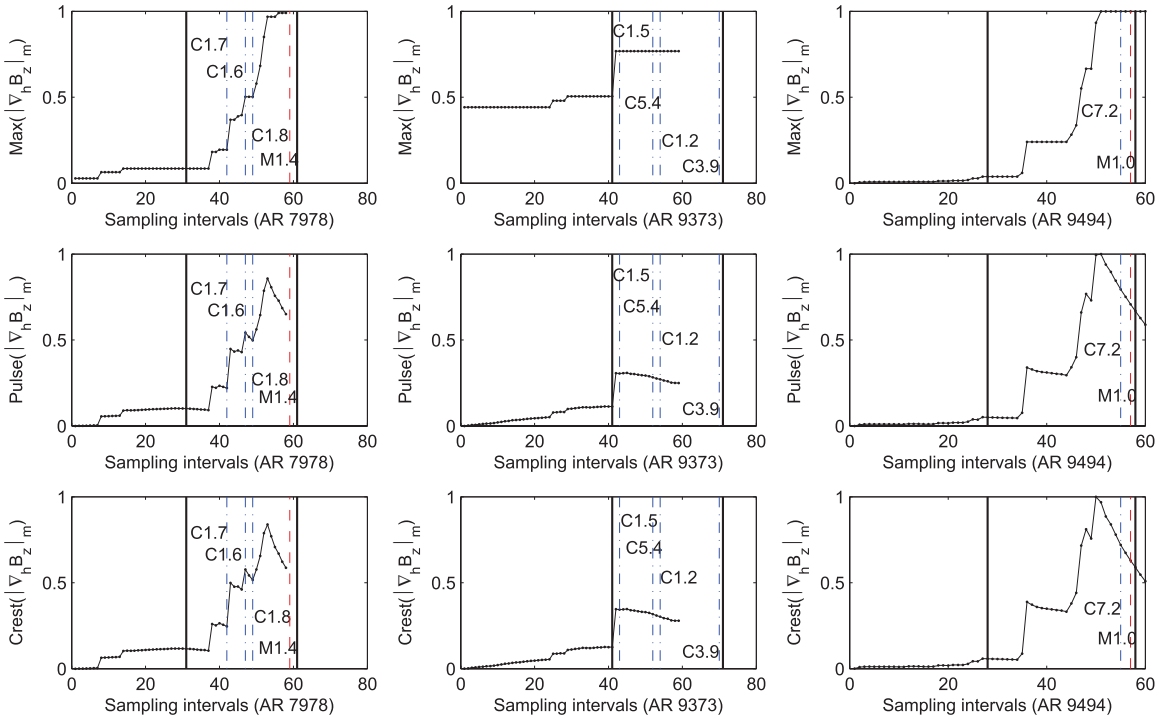

Figure 1. Sensitive changes for a single measurement in AR 7978, AR 9373, and AR 9494 (left to right). The sampling interval is 96 minutes. The first vertical solid line indicates where I_{tot} crosses the specified threshold of 10. The forecasting time is indicated by the second vertical solid line. The dash-dotted lines indicate the eruption of C level flare, and the dashed lines indicate the eruption of M level flare. The corresponding flare level is labeled beside these vertical lines.
(A color version of this figure is available in the online journal.)

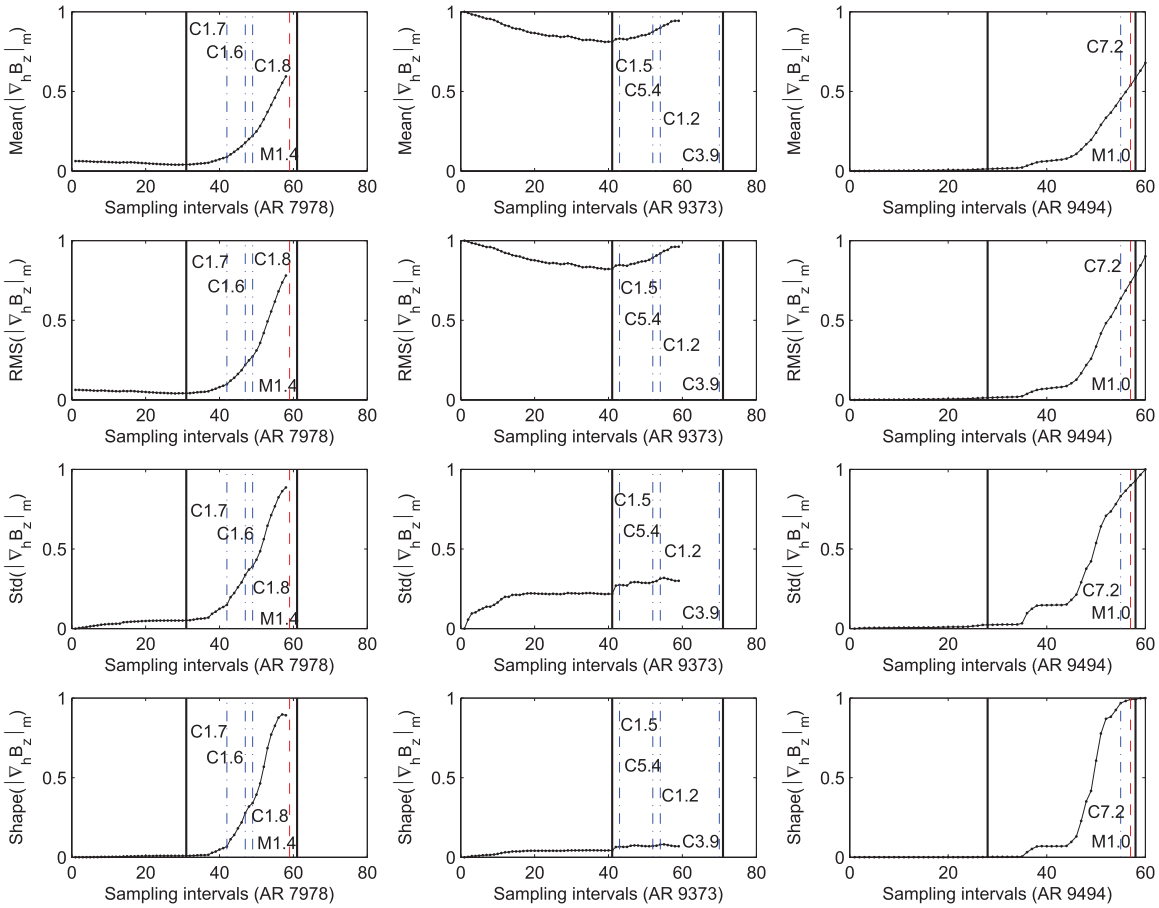

Figure 2. Robust changes for a single measurement in AR 7978, AR 9373, and AR 9494 (left to right). The sampling interval is 96 minutes. The definitions of the vertical lines are the same as that in Figure 1.

(A color version of this figure is available in the online journal.)

The Bayesian network represents the conditional independent relationships of $P(\mathbf{X}|y)$ as a directed acyclic graph together with the corresponding conditional probability tables. The nodes in the network represent predictors and the decision. The arcs between nodes represent the probabilistic dependant relationships, that is, the probability of a node is only dependent on its parent nodes as shown as follows:

$$\prod_{i=1}^{n} P(x_i|y, x_1, x_2, \ldots, x_{i-1}) = \prod_{i=1}^{n} P(x_i|Pa_i), \tag{7}$$

where Pa_i is the parent nodes of x_i. These dependencies are used to simplify the estimation of the probability distribution. The probability values of each node are represented in the conditional probability table.

Generally speaking, the Bayesian network consists of the probabilistic network structure and the conditional probability tables. The probabilistic network structure represents the qualitative-dependent relationships among the predictors, and the conditional probability tables determine the quantitative dependence between the node and its parents.

4.1. Learning Probabilistic Network Structure

The structure of the Bayesian network is learned from data to represent the probabilistic dependences among the predictors. We wish to learn the network structure that is most likely to reflect the relationships between predictors in the data set. This can be stated as follows:

$$\begin{aligned} B_{S_{\max}} &= \arg\max_{B_S}(P(B_S|D)) \\ &= \arg\max_{B_S}(P(D|B_S)P(B_S)), \end{aligned} \tag{8}$$

where D is the observed data, B_S stands for a network structure, and $B_{S_{\max}}$ is the B_S that maximizes the $P(B_S|D)$.

In order to efficiently compute $P(D|B_S)P(B_S)$ in Equation (8), the following four assumptions are introduced (Cooper & Herskovits 1992). (1) The predictors are discrete. (2) Given a network, the samples occur independently. (3) The predictors are not missing values. (4) Before observing the data set, we are indifferent to the numerical probabilities to place on the network structure. Under these assumptions, we obtain the following formula (Cooper & Herskovits 1992):

$$P(D|B_S)P(B_S) = P(B_S)\prod_{i=0}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(r_i-1+N_{ij})!}\prod_{k=1}^{r_i}N_{ijk}!, \tag{9}$$

where $P(B_S)$ is the prior probability of B_S. r_i is the number of possible values for a predictor x_i in the network B_S. Each predictor x_i in B_S has a set of parents $Pa(x_i)$, q_i is the number of different values to which $Pa(x_i)$ can be instantiated. n is the number of predictors in the data set D. N_{ijk} is the number of cases in data set D in which predictor x_i has the kth value and its jth parent node is selected. $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$.

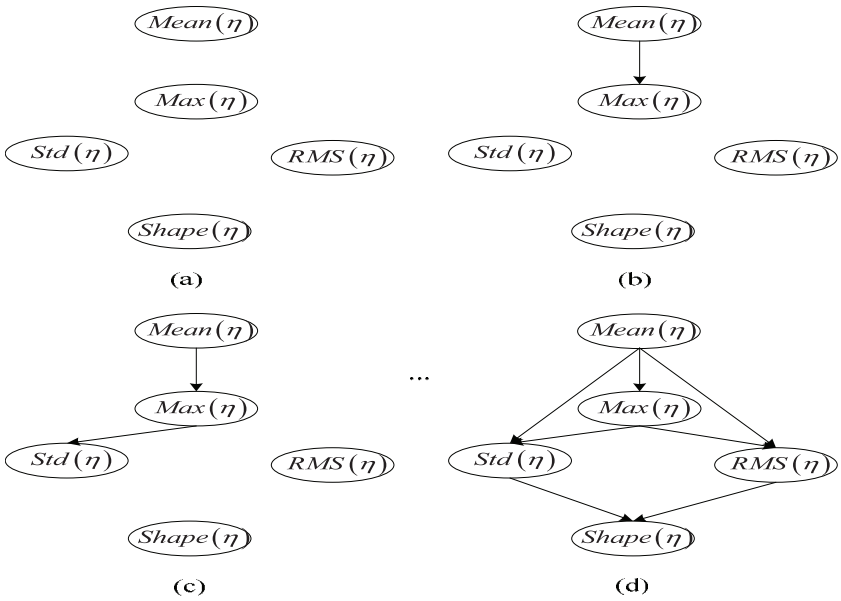

Figure 3. Process of network generation. Algorithm starts with the network having no arcs (a). At each stage, arc that most dramatically increases the probability of the network is added ((b) and (c)). When there is no node that could be added, this process is finished (d).

It is computationally intractable that finding the best B_S in all the combinations of predictors, so the heuristic search strategy is used to find a proper network structure. We can begin with the unconnected network as shown in Figure 3(a). At each stage, the arc that results in the best structure is added. The respective metrics which determines how adequately a probability distribution captures the dependencies between predictors in the data set is defined as

$$\text{Score}_{K2}(B_S, D) = P(B_S, D) = P(D|B_S)P(B_S). \quad (10)$$

The process of adding arcs recurs, until the stop criterion is reached. It is shown in Figures 3(b)–3(d). Finally, an approximate solution of $B_{S_{\max}}$ is found.

Theoretically, a best-fitted structure of the Bayesian network can be found by the K2 metrics. However, the number of possible structures exponentially grows with the number of nodes in the Bayesian network. It is hard to exhaustively enumerate all the structures, so a heuristic search method is used to find the proper structure, but it is not guaranteed that the searched structure is the global optimization. Greedy search algorithm is one of the heuristic search methods. It makes the locally optimal choice at each stage to approach the global optimum solution. When there are too many nodes in the Bayesian network, this search algorithm is applied to find a proper network structure quickly. For each node, the search procedure is used to find its set of parent nodes with iteratively adding a parent node with the largest increases of the K2 score, until the number of parent nodes exceeds the settled value or the K2 score is not increased any more. Given that all the input parameters are independent, the Bayesian network is simplified as the naive Bayes model. When a suitable network structure is learned, the conditional probability tables of nodes in the Bayesian network should be estimated.

4.2. Learning Conditional Probability Tables

Conditional probability tables reflect the quantitative relationships between a node and its parents in the Bayesian network. The conditional probability of a node is determined by its parent nodes only. For example, the conditional probability table of node Mean(η) in Figure 4 is calculated as $P(\text{Mean}(\eta)|pa(\text{Mean}(\eta))) = P(\text{Mean}(\eta)|\text{flare})$. The continuous variables can be discretized by an entropy-based discretization method (Jin et al. 2009). This method does not simply divide the continuous predictor into equal-width or equal-frequency bins. The bins with proper width are selected to maximize the flare information provided by the predictor which has been discretized with these bins. Each value in the conditional probability table is estimated by the Bayesian method. Bayesian parameter estimation provides a distribution of the parameter. Usually, the mathematical expectation of the parameter ($E(\theta_{ijk}|D) = \frac{N_{ijk}+1}{N_{ij}+r_i}$) is used to estimate the real value of the parameter. So, the conditional probabilities are estimated as follows (Cooper & Herskovits 1992):

$$P(x_i = k|pa(x_i) = j) = \frac{N_{ijk} + 1}{N_{ij} + r_i}. \quad (11)$$

Once the conditional probability tables of each nodes are calculated, a complete Bayesian network is built. It can be used to capture the probabilistic relationships among predictors and automatically construct a probabilistic expert system.

4.3. Bayesian Networks for Flare Level Prediction

Two Bayesian networks are trained by the feature extracted data (BN_F) and the raw sequential data (BN_R), respectively.

The Bayesian network trained by the feature extracted data is shown in Figure 4. Only the features of the sequence η are drawn in this figure to clearly reflect the relationships among the features. Some reasonable conclusions are given: the standard deviation of the sequence is influenced by its maximum and mean. The root mean square of the sequence is influenced by its mean and maximum. The pulse factor of the sequence ($\frac{\text{Max}}{\text{Mean}}$) is influenced by the shape factor ($\frac{\text{rms}}{\text{Mean}}$) and the crest factor ($\frac{\text{Max}}{\text{rms}}$). However, the structure of the Bayesian network is not the global optimum. For example, the pulse factor is not directly influenced by the mean, except in that the mean enters into the shape factor

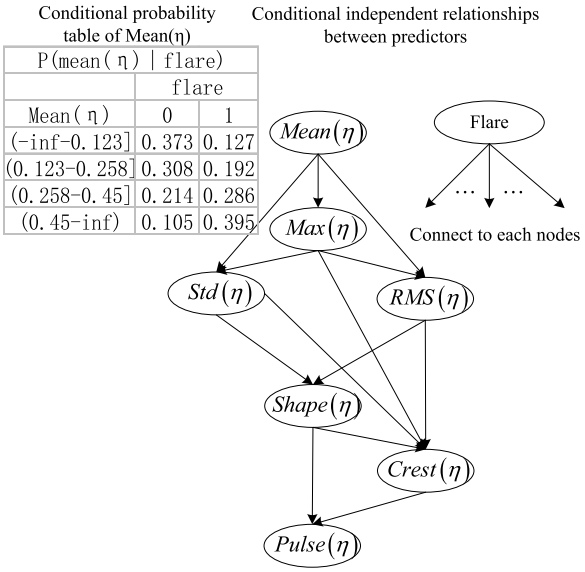

Figure 4. Bayesian network model of solar flare level prediction trained by feature extracted data. Only network structure of features of η and conditional probability table of node Mean(η) are drawn.

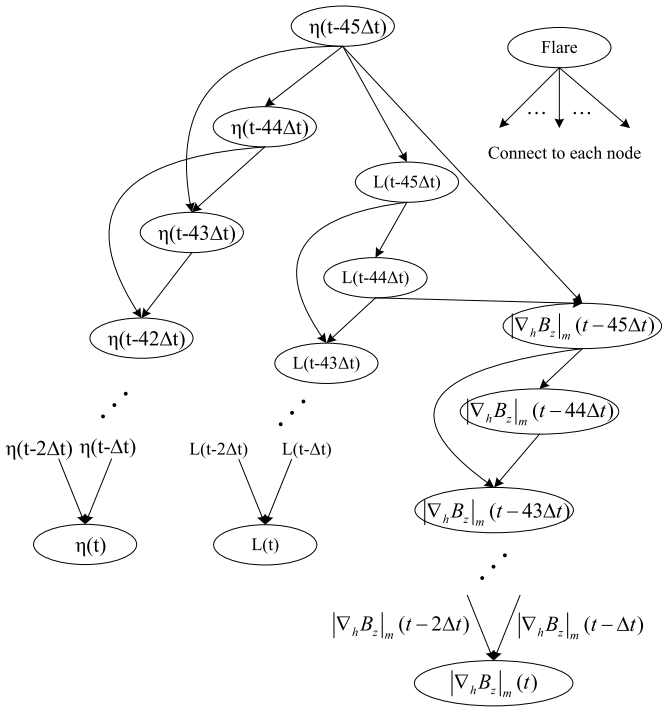

Figure 5. Bayesian network model of solar flare level prediction trained by raw sequential data.

which influences the pulse factor. The heuristic method is used to search the proper network structure to avoid the exponential growth of the number of possible structures with the increase of the number of nodes. For a greedy algorithm, the searched network structure is just an approximation of the global optimal solution.

The Bayesian network learned from the raw sequential data is shown in Figure 5, and two main conclusions are given as follows:

1. Most predictors are dependent on their previous two predictors. For example, given $\eta(t-2\Delta t)$ and $\eta(t-\Delta t)$, $\eta(t)$ is independent of other predictors.
2. $L(t-45\Delta t)$ is influenced by $\eta(t-45\Delta t)$, and $|\nabla_h B_z|_m$ $(t-45\Delta t)$ is influenced by $\eta(t-45\Delta t)$ and $L(t-44\Delta t)$. In fact, the influences among the parameters are more generally true at other times. However, the Bayesian network algorithm encodes these influences at $t-45\Delta t$ because it is not a global optimal algorithm. This means that not all the influences can be expressed exactly. The relationships among the predictors expressed by the learned Bayesian network are consistent with the physical explanations in Wang et al. (2009). From a geometrical point of view, the number of singular points describes the topological complexity of photospheric magnetic fields with the surface scale, the length of neutral line is the line scale, and the maximum horizontal gradient is a point function with the local scale. The measure of the line scale is influenced by the measure of the surface scale, and the measure of the local scale is influenced by the measure of the surface scale and line scale simultaneously.

5. EXPERIMENTAL RESULTS AND ANALYSES

5.1. Performance Evaluation

We treat solar flare level prediction as a binary classification task. The flaring sample is considered as positive class, and the non-flaring sample is considered as negative class. Therefore, there are four possible outcomes shown in Table 3. As shown in Table 3, samples correctly classified as "Positive" are defined as true positive (TP), while the samples correctly classified as "Negative" are defined as true negative (TN). On the other hand, samples wrongly predicted as "Positive" are defined as false positive (FP), and samples wrongly predicted as "Negative" are defined as false negative (FN). Prediction performance can be measured using TP rate and TN rate.

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

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:

$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

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:

$$\text{TN rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{13}$$

TP rate and TN rate are used to evaluate the flaring and non-flaring accuracy, respectively. When the occurrence of events is very rare, TP rate and TN rate could avoid the drawback of success rate. Taking into account the frequency of events, the Heidke skill score (HSS) defined in Equation (14) is used to quantify the performance of a forecasting method (Balch 2008).

$$\text{HSS} = \frac{\frac{\text{TP+TN}}{N} - \frac{(\text{TP+FP})(\text{TP+FN})+(\text{TN+FN})(\text{TN+FP})}{N^2}}{1 - \frac{(\text{TP+FP})(\text{TP+FN})+(\text{TN+FN})(\text{TN+FP})}{N^2}}, \tag{14}$$

where N is the total number of samples in the testing set.

Table 4
Values in Contingency Table for the Testing Process of NB_R, NB_F, BN_R, and BN_F

Number of Samples	NB_R		NB_F		BN_R		BN_F	
	PP	PN	PP	PN	PP	PN	PP	PN
AP	604 ± 7	248 ± 7	552 ± 8	299 ± 8	725 ± 4	126 ± 4	726 ± 8	125 ± 8
AN	472 ± 6	1793 ± 6	434 ± 6	1832 ± 6	351 ± 8	1915 ± 8	283 ± 9	1982 ± 9

Notes. 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 NB_R, NB_F, BN_R, and BN_F

Performance Evaluation	NB_R	NB_F	BN_R	BN_F
TP rate (%)	70.9 ± 0.8	64.9 ± 0.9	85.2 ± 0.5	85.3 ± 0.9
TN rate (%)	79.2 ± 0.3	80.9 ± 0.3	84.5 ± 0.4	87.5 ± 0.4
HSS	0.463 ± 0.007	0.436 ± 0.006	0.644 ± 0.006	0.688 ± 0.011

Note. Uncertainty of performances is estimated by standard deviation.

HSS can range from -1 to $+1$, where -1 stands for no correct predictions, $+1$ stands for all correct predictions, and 0 indicates that the predictions have been generated mainly by chance.

5.2. Performance of Bayesian Network Models

The present model is applied to analyze sampling intervals for regions which produced at least a C1.0 flare within the 48 hr window following the sampling interval. The data set that contained 8510 flaring samples and 22654 non-flaring samples from 1996 April 15 to 2004 January 10 is used to build the solar flare level prediction model. The number of non-flaring samples is more than the number of flaring samples in the data set. This is the class imbalance problem in the data mining community (Japkowicz & Stephen 2002). The model learned from the imbalanced data set could bias the majority class in the data set. However, we generally paid more attention to the samples in the minority class, so undersample technique is used to overcome this problem. In the undersample process, the training set is randomly undersampled until it is balanced, and then the model is learned from the balanced training data set. Finally, the performance of the model is evaluated using the imbalanced testing set. The data set is divided into tenfolds, therein ninefolds are used for training and the remaining onefold for testing. This process is repeated 10 times, and the average value of test accuracies is considered as the estimation of the prediction performance. The uncertainty of the performance is evaluated by the standard deviation of 10 times experiments. The algorithms of the Bayesian network are implemented in Waikato Environment for Knowledge Analysis (WEKA) which is a data mining software in JAVA (Witten & Frank 2005). The testing values of the contingency table are shown in Table 4.

The performances of the Bayesian networks model trained with the raw sequential data (BN_R) or the feature extracted data (BN_F) and the naive Bayes model trained with the raw sequential data (NB_R) or the feature extracted data (NB_F) are listed in Table 5. In order to present the influence of the learned dependent relationships on the flare level prediction, the performances between the Bayesian network and the naive Bayes model which assumes that all the predictors are conditionally independent are compared. The performance of the Bayesian network model is higher than the performance of the naive Bayes model. Because the Bayesian network model relaxes the strong assumption of the naive Bayes model, the performances of two Bayesian network models are improved. In order to reduce the dimension of the sequential predictors, for each type of measurement, seven features are extracted to reflect the variation of the raw sequences. The performance of BN_F is slightly higher than the performance of BN_R. Furthermore, the feature extraction based method with the lower dimension is more rapid to learn and forecast, and its network structure is more compact.

Table 6
Performance Comparisons of LVQ, C4.5, BN_R, and BN_F

Performance Evaluation	LVQ	C4.5	BN_R	BN_F
TP rate (%)	82.6	81.7	85.2	85.3
TN rate (%)	84.1	83.4	84.5	87.5
Number of nodes	Black-box	1899	139	22
Comprehensibility	Bad	Medium	Good	Excellent

5.3. Comparisons with Other Methods

Many solar flare forecasting approaches have been proposed (Barnes & Leka 2008; Colak & Qahwaji 2009). However, it is difficult to quantitatively compare the performance of this method with the performance of other methods, because of the different definitions of the flare level and the selection criteria of active regions.

With the same definition of the flare level, the threshold, the forecasting time, and the selection criteria of active regions, this method is quantitatively compared with the methods proposed by Yu et al. (2009). As shown in Table 6, the performance of Bayesian network models is compared according to both their accuracies and the comprehensibility. The accuracies of BN_R and BN_F are slightly higher than the accuracies of the LVQ and C4.5 decision tree. Furthermore, the Bayesian network models are more understandable than the LVQ and C4.5 decision tree. More importantly, the models of the decision tree and Bayesian network can be presented as a graph, so we can quantitatively compare their comprehensibility. The complexity of these models can be measured by their number of nodes,

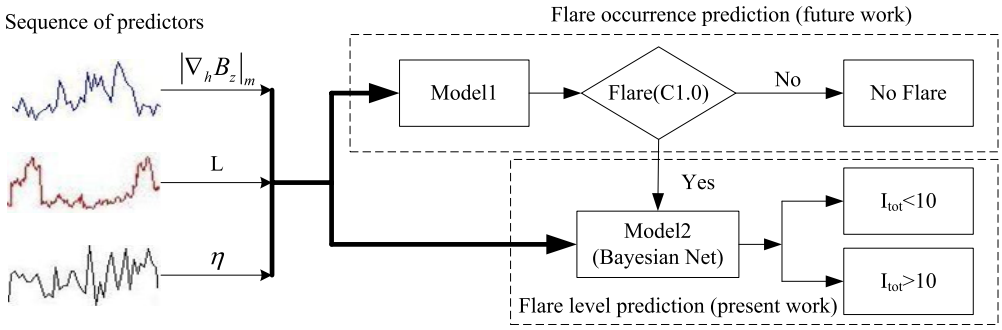

Figure 6. Schematic diagram of whole prediction for short-term solar flare prediction.
(A color version of this figure is available in the online journal.)

and the more complex the model is, the less comprehensible the model is. As shown in Table 6, the LVQ network is a black-box model whose results cannot be understood. According to information theory, each predictor of the C4.5 decision tree is divided into several intervals, and each node represents the selected predictor divided the samples into different subsets by the certain interval. Each predictor may be used several times with their different intervals, that is, there are lots of nodes in the decision tree model. For a complex problem, although the C4.5 decision tree can be transformed to comprehensible if-then rules, the scale of the rules obtained from the decision tree is too large to understand. The comprehensible information of the Bayesian network model is represented by its structure. The number of nodes in the structure is equal to the number of predictors used in this model. So, the relationships among predictors are more clear in the Bayesian network. The number of nodes of C4.5 decision tree is 10 times more than the number of nodes of BN_R, and 80 times more than the number of nodes of BN_F. So the comprehensibility of Bayesian network models of solar flare level prediction is greatly improved over the C4.5 decision tree model of solar flare level prediction. Furthermore, the rules generated by the C4.5 decision tree are determinate, however, the relationships between the predictors and the flare are not determinate. The Bayesian network is an approach for reasoning under uncertainty. It can capture the probabilistic relationships among predictors and these relationships are explicable. So it is reasonable to build the short-term solar flare level prediction model using the Bayesian network approach.

6. CONCLUSIONS AND DISCUSSION

The uncertainty between magnetic field parameters and flare urges the Bayesian network to be used to build the probabilistic reasoning model. The Bayesian network approach encodes the conditional independent relationships among predictors. Comparing the performance of the Bayesian network model with the naive Bayes model, the importance of these relationships to flare level prediction is shown, and the relationships of predictors are consistent with the physical explanations in Wang et al. (2009). So, this implies that the physical rules can be obtained from the observational records, and the understandability of the Bayesian network can help researchers to analyze the problem domain better.

The dimension is reduced by extracting the features from the sequences. Meanwhile, it provides a method to construct new predictors. The accuracies of the Bayesian network models trained with the raw sequential data and the feature extracted data are slightly higher than the accuracies of the methods proposed by Yu et al. (2009). Because of its uncertainty reasoning ability and comprehensibility, the Bayesian network is recommended to be used to forecast flares.

In the future, more features better reflecting the physical nature should be extracted from the observational sequences, and this method can be improved by considering the previous flare records like the method of Wheatland (2004). As the Qahwaji & Colak (2007) system did, the whole solar flare prediction system shown in Figure 6 will be built by adding the module of flare occurrence prediction.

This work is supported by the National Basic Research Program of China (973 Program) through grant no. 2006CB806307 and the National Natural Science Foundation of China (NSFC) through grant nos. 10673017, 10733020, and 10978011. We thank the *SOHO*/MDI consortium for the data. *SOHO* is a project of international cooperation between ESA and NASA. This paper has benefited from the comments of the anonymous reviewer.

REFERENCES

Balch, C. C. 2008, Space Weather, 6, S01001
Barnes, G., & Leka, K. D. 2008, ApJ, 688, L107
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
Colak, T., & Qahwaji, R. 2008, Sol. Phys., 248, 277
Colak, T., & Qahwaji, R. 2009, Space Weather, 7, S06001
Cooper, G. F., & Herskovits, E. 1992, Mach. Learn., 9, 309
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
Gallagher, P. T., Moon, Y. J., & Wang, H. 2002, Sol. Phys., 209, 171
Georgoulis, M. K., & Rust, D. M. 2007, ApJ, 661, 109
Japkowicz, N., & Stephen, S. 2002, Intell. Data Anal., 6, 429
Jin, R., Breitbart, Y., & Muoh, C. 2009, Knowl. Inf. Syst., 19, 1
Jing, J., Song, H., Abramenko, V., Tan, C., & Wang, H. 2006, ApJ, 644, 1273
Koskinen, H., Eliasson, L., Holback, B., Andersson, L., Eriksson, A., & Mälkki, A. 1999, SPEE Final Report, FMI Reports 1999:4, Finnish Meteorological Institute, Helsinki
Leka, K. D., & Barnes, G. 2003a, ApJ, 595, 1277
Leka, K. D., & Barnes, G. 2003b, ApJ, 595, 1296
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
Mitchell, T. M. 1997, Machine Learning (New York: McGraw-Hill)
Qahwaji, R., & Colak, T. 2007, Sol. Phys., 241, 195

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. 2001, Sol. Phys., 203, 87
Wheatland, M. S. 2004, ApJ, 609, 1134
Wheatland, M. S. 2005, Space Weather, 3, S07003
Witten, I. H., & Frank, E. 2005, Data Mining: Practical Machine Learning Tools and Techniques (San Francisco, CA: Morgan Kaufmann), 271
Yu, D. R., Huang, X., Hu, Q. H., Zhou, R., Wang, H. N., & Cui, Y. M. 2010, ApJ, 709, 321
Yu, D. R., Huang, X., Wang, H. N., & Cui, Y. M. 2009, Sol. Phys., 255, 91