Short-term solar flare forecast

Victor Chernyshov, Dmitry Laptev, Dmitry Vetrov
Department of Computational Mathematics and Cybernetics
Moscow State University, Moscow, Russia
webcreator18@gmail.com, laptev.d.a@gmail.com, vetrovd@yandex.ru

Abstract

In this paper a new automated hybrid method for short-term flare forecasting is introduced and suggested for future use.

At the initial stage we created a flare base, and an image base for 1996–2009 years interval.

Further, we derived simple and efficient parametric precedent model, which turned our prediction problem into two-class classification problem, and developed machine learning-based procedures for features extraction from both magnetograms and continuum images.

We develop an experimental protocol to estimate the accuracy of obtained decision rules and report 63% to 82% on balanced data (maximum worst-case), 73% to 90% on real data depending on the choice of precedent model configuration.

Keywords: *solar flare forecast, precedent model, spherical correction, sunspots clustering, magnetogram segmentation, active regions localization.*

1. INTRODUCTION

A *solar flare* is a sudden brightening observed over the Sun surface or the solar limb, which is interpreted as a large energy release. *Sunspots* are temporary phenomena on the photosphere of the Sun that appear in visible spectrum as dark spots compared to surrounding regions. An ***a**ctive **r**egion (AR)* on the Sun is an area with an especially strong magnetic field. Sunspots usually form in active regions.

X-rays emitted by solar flares can affect Earth's ionosphere and disrupt long-range radio communications and disturb operation of navigation systems. The most violent eruptions may affect satellite or cause problems with power grid.

Solar flares research has shown that X-ray flares are closely related to sunspots and active regions [1]. So, a number of flare forecasting methods based on this relationship has been proposed. McIntosh [2] revised sunspot classification and specially dedicated system called Theophrastus was developed in 1987. The method depends on human expert.

In 2009 Qahwaji and Colak presented an automated hybrid computer platform for the short-term prediction of significant solar flares using Solar and Heliospheric Observatory (SOHO)/Michelson Doppler Imager (MDI) images[3]. Proposed method incorporates sunspot grouping (both MDI continuum and magnetogram images are used), McIntosh-based classification, and, afterwards, flare prediction using neural networks.

Yu et al. [4] analyzed the influence of sequences of magnetic-based parameters on the flare level and proposed bayesian solar flare prediction models.

Very recently Falconer et al. introduced their tool for empirical forecasting of major flares from a proxy of active region free magnetic energy [5]. Their method is mainly focused on measuring the proxy of the active regions free magnetic energy, and the empirical relationship is then used to convert the free magnetic energy proxy into an expected event rate.

In our research we included features of both types (magnetic and continuum), derived adequate precedent model in order to use Support Vector Machine (SVM) classification algorithm. As a result, fully-automated testing system was built and good short-term prediction results achieved.

The rest of the paper is organized as follows. The solar data used in this paper are described in Section 2.. A precedent model is proposed in Section 3.. Features extraction is discussed in Section 4.. The implementation of the system and testing results are reported in Section 5..

2. DATA

In this study we used data from the publicly available NOAA solar flare catalogue[1] and images in FITS format from SOHO/MDI in the resolution 1024px x 1024px (can be downloaded via web interface[2]). SOHO/MDI provides both continuum ("white-light", 2–5 per day) and magnetogram observations (in the vicinity of the Ni I 6767.8 Å photospheric absorption line, 6–14 per day) of the Sun.

We created a flare base, consisting of 24658 events, and an image base, including 40573 magnetograms and 14927 continuum images from 1996–2009 years interval. All flares are divided into flare series (using the NOAA active region number): 1397 series, maximum flares in series is equal to 154, minimum — 1, in average — 8. We use notation $(i_f,\ j_f)$, where i_f — series number, j_f — flare number in the series.

3. PRECEDENT MODEL

Solar flares are classified as A, B, C, M or X according to the peak flux as measured on the GOES spacecraft. Hereinafter we use the following notation: *strong flares* are flares not not weaker than M_f, *weak flares*, respectively, are weaker than M_f. The exact values of parameters we used are given in the end of the section. It is assumed that precedent model is built on the training stage of the method, so, we know when and where a flare occurred.

Stage 0. For simplicity, let's fix flare F, which will correspond to a precedent. It uniquely localizes an active region on every image within flare's prehistory. We consider only images within flare's T_{ph} days prehistory. Hereinafter the words "preceding", "closely located", "nearest" should be treated in the context of time.

Stage 1: base precedent. We choose a magnetogram image and the nearest preceding magnetogram image located not closer than Δ_f days. Denote the first image as "head magnetogram". For the head image we determine the most closely located continuum image ("head continuum"), for which we also find the nearest preceding image located not closer than Δ_f days. Two pairs of images and active region on them we denote as *base precedent*.

Stage 2: positive and negative precedents. *Positive precedent*

[1] http://www.ngdc.noaa.gov/stp/solar/solarflares.html
[2] http://sohowww.nascom.nasa.gov/data/archive/

(class 1) is a base precedent, which meets following requirements:
1. Time from a head magnetogram to the flare does not exceed T_f.
2. Flare strength is not less than M_f.
3. The nearest flare, which meets 1. and 2., corresponds to the head magnetogram.
Negative precedent (class 0) is a base precedent, which is not a positive one.

T_f, M_f, Δ_f, T_{ph} are structural parameters of the precedent model. In our research we suppose $M_f \in M_{set} = \{C5.0, M1.0, M5.0, X1.0\}$, $T_f \in T_{set} = \{1.0, 1.5, 2.0\}$ days, $\Delta_f = 0.25$ days, $T_{ph} = 4.0$ days, so, we investigate different configurations.

4. FEATURES

A general scheme for features extracting is the following:
1. We fix a head magnetogram and build a precedent according the model.
2. Sunspot groups are localized on the nearest to the flare cont-image.
3. If there is a no correspondence between the flare and one of the sunspot groups, we start building another precedent. Otherwise we have the sunspot group and all its characteristics (including bounding box parameters).
4. Steps 2–3 are performed for the second cont-image. After that we are able to extract cont-features from the pair of cont-images (Section 4.1).
5. We apply tracking procedure to the found sunspot group (to be precise, to the center of its bounding box) to locate active regions on both magnetograms and extract from them magn-features (Section 4.2).

The described above scheme is applied for the whole image base for every configuration $(M_f, T_f) \in \{M_{set} \times T_{set}\}$. As a result we have $|M_{set}| \cdot |T_{set}|$ datasets. We use a notation $X_{M_f, T_f} \in \mathbf{R}^{m \times n}$ for features table, $Y_{M_f, T_f} \in \{0, 1\}^{m \times 1}$ for class labels, where m — number of features, n — number of precedents.

4.1 Continuum-based features

4.1.1 Continuum images preprocessing

If we have plane projection of a semisphere, we can not avoid distortions: the same areas on a semisphere in general case are not the same after projection. A point on a semisphere is denoted as $P'(x',\, y',\, z')$, it corresponds to a point $P'_{xy}(x',\, y')$ on the visible image. Supposed that spherical correction maps a point P'_{xy} to a pixel $P(x,\, y)$ on spherically corrected image.

Assuming that proposed mapping keeps ratio 1

$$\frac{\beta}{\frac{\pi}{2}} = \frac{\sqrt{x^2 + y^2}}{R} \tag{1}$$

$$\alpha = \arctan(\frac{y}{x}), \rho = r \sin\beta,\; x' = \rho\cos\alpha,\; y' = \rho\sin\alpha,$$

ρ is the length of vector $O'P'$ projection on plain $O'XY$,

$\beta = \widehat{ZO'P'}$, $O'Z$ is “eye-pointed” axis.

we get coordinates of P':

$$\begin{cases} x' = sgn(x) r \cos(\arctan(\frac{y}{x})) \sin(\frac{\pi}{2} \frac{\sqrt{x^2+y^2}}{R}), \text{if } x \neq 0 \\ y' = sgn(x) r \sin(\arctan(\frac{y}{x})) \sin(\frac{\pi}{2} \frac{\sqrt{x^2+y^2}}{R}), \text{if } x \neq 0 \\ x' = 0,\; y' = 0, \text{if } x = y = 0 \\ z' = \sqrt{r^2 - (x')^2 - (y')^2} \end{cases}$$

To get approximation of the intensity in $P'(x',\, y')$, we use bilinear interpolation.

Further a term ”image” is used for images with applied spherical correction procedure.

Before the preprocessing we create an average background cont-image (pixel-by-pixel median through 500 cont-images); it is also called ”quiet Sun image”.

We also perform Gaussian blurring with $\sigma = 0.7$ and $window\ size = 7$ and cut-off beyond $\alpha_{cut} = 65^\circ$ (heliocentric degrees) to get rid of limb darkening and overblurring effects.

4.1.2 Adaptive binarization

We worked out simple adaptive binarization procedure to localize sunspot groups in the Sun (see figure 1 c): we set knowingly high threshold $T_0 = 0.98$ and start decreasing it with the step $\tau = 0.002$. At every step binarizarion is performed, thus, we know the number of connected components and their sizes. If connected components number doesn't exceed $N_{max\_comp} = 100$, and among them there are no components with area is more than $0.12 * (R * \sin\alpha_{cut})^2)$, then we have found appropriate binarization threshold T_{bin}. Exceeding maximum loop iterations number $N_{max\_iter} = 80$ is an alternative loop exit condition.

4.1.3 Umbra and penumbra segmentation

When viewed through a telescope, sunspots have a dark central region known as the *umbra*, surrounded by a somewhat lighter region called the *penumbra*. Umbra and penumbra characteristics can be used for solar flare prediction [2].

Area of the sunspot is denoted as S_F. Quiet Sun intensity value I_q is a non-zero intensity, corresponding to the peak on the histogram of a preprocessed image (we used a partition of $[0, 1]$ in $N_{hist\_bins} = 1000$ equal intervals). Our approach to umbra and penumbra segmentation is based on the method proposed by Zharkov et al. [6], which incorporates umbra (T_u) and penumbra (T_p) thresholds:
1. if $S_F \leq 5$ pixels then assign the thresholds: $T_p = 0.91 * I_q$; $T_u = 0.6 * I_q$
2. if $S_F > 5$ pixels then assign the thresholds: $T_p = 0.93 * I_q$; $T_u = \max(0.55 * I_q, \mathbb{E}P - \Delta P)$, where $\mathbb{E}P$ is the mean intensity value, ΔP is the standard deviation for pixel intensities in the region F.

Our modification uses only T_p threshold. Thus, sunspot pixels, which are not in umbra, are supposed to be in penumbra; due rather flexible adaptive binarization, T_{bin} could be considered as T_p analogue (see figure 1 e).

4.1.4 Sunspots clustering

After localizing sunspots we need to combine them into several sunspot groups. For this purpose we use agglomerative hierarchical clustering procedure with euclidian metrics and Ward linkage (see figure 1 f).

4.1.5 Extracted features

Described above stages of cont-images processing make calculating cont-features for a pair of images possible. Cont-features are calculated using head continuum image, their speed of change — using the both images. A sunspot group is unambiguous defined by the solar flare, corresponding to a precedent.

Extracted cont-features:
1. umbra square of the sunspot group,
2. penumbra square of the sunspot group,
3. speed of change for 1–2 (px/sec).

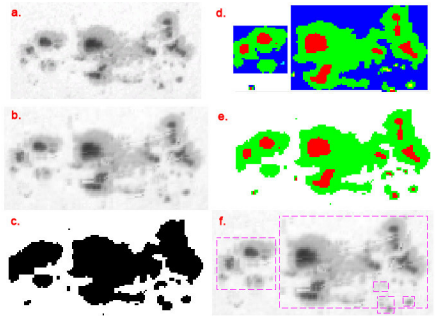

Figure 1: Vicinity of the sunspot group, localized on cont-image 4240.0003, at different stages of features extraction: (a) — initial, (b) — after spherical correction, (c) — the result of adaptive binarization, (d) — segmented umbra and penumbra (using Zharkov et. al method), (e) — segmented umbra and penumbra using our method, (f) — after clustering all neighboring sunspots form a sunspot group.

4.2 Magnetic-based features

A magnetogram image is a representation of the variations in strength of a magnetic field. Black regions correspond to strong positive magnetic field, white regions — to strong negative magnetic field, neutral Sun regions are marked with grey color. We extract magn-features in three stages: performing segmentation of a magnetogram, finding active regions, counting features.

4.2.1 Segmentation

The first step is to find the areas of strong positive magnetic fields, strong negative magnetic fields and neutral areas. Proposed method is based on variational approximation [7] with the use of global constraints [8], which give us a possibility to include some physically-driven conditions in our model (i. e. the equality of positive and negative fluxes within AR).

Let i be the pixel of an image with the value of the magnetic field I_i, N — number of pixels, Z_i — class label for pixel i, $\mathcal{E}$ — neighborhood system.

In these terms the discrete optimization task could be formulated as

$$p(I|Z) \propto \prod_{i=1}^{N} \varphi_i(Z_i) \prod_{j \in \mathcal{E}(i)} \phi(Z_i, Z_j) \to \max_Z \qquad (2)$$

Where $\phi(Z_i, Z_j) = e^{C_{pair}[Z_i \neq Z_j]}$ — pairwise term, $\varphi_i(Z_i)$ — unary term: $\varphi_i(1) = e^{-C_1\sqrt{|2000-I_i|}}$, $\varphi_i(2) = e^{-C_1\sqrt{|2000+I_i|}}$, $\varphi_i(3) = e^{-C_2|I_i|}$.

This task is solved with factorized approximation

$$p(Z|I) \approx q(Z) = \prod_{i=1}^{N} q_i(Z_i) \qquad (3)$$

Minimizing KL-divergence between $q(Z)$ and $p(Z|I)$ the following equation could be obtained:

$$q_j^*(Z_j) = \frac{\exp(\mathrm{E}_{i \neq j} \log p(Z)}{\int \exp(\mathrm{E}_{i \neq j} \log p(Z) dZ_j}$$

Using the exact form of $p(Z)$ we obtained iterative process

$$q_i^{new}(Z_i) = \frac{1}{C} \exp\left(\log(\varphi_i(Z_i)) - C_{pair} \sum_{t \in \mathcal{E}(i)} \sum_{j \neq i} q_j^{old}(Z_j) \right)$$

4.2.2 Active region search

For localizing an active region on the Sun we use the method introduced in [9], which is based on the branch and bounds approach to maximizing the functional defined on a rectangle. The initial bounding rectangle on a magn-image is obtained through tracking and the following enlarging it in 2.5 times.

Let R be the rectangle, $A_i = q_i(1) + q_i(2)$ and $B_i = q_i(3)$. The active region could be found by maximizing the following functional:

$$F(R) = \alpha \sum_{i \in R} A_i - \sum_{i \in R} B_i + \beta \sqrt{Area(R)} \sum_{i \in border\ of\ R} B_i.$$

The global optimum could be found using the function $\hat{F}(\mathbf{R})$ which is the top border of $F(R)$ and equal $F(R)$if $\mathbf{R} = \{R\}$. The optimization procedure is then performed as follows: at the first step $\mathbf{R}$ consists of all the possible rectangles R, the sets $\mathbf{R}$ are placed in the priority queue in a decreasing order of $\hat{F}(\mathbf{R})$, in every step the first element of the queue is divided into two and returned back to the queue. If the first element of the queue consists of just one rectangle, that is the answer.

In our case the function $\hat{F}(\mathbf{R})$ could be chosen as follows:

$$\hat{F}(\mathbf{R}) = \alpha \sum_{i \in R_{small}} A_i - \sum_{i \in R_{big}} B_i + \beta \sqrt{Area(R_{big})}$$
$$\Big(\max_{R \in \mathbf{R}} \sum_{i \in left\ border\ of\ R} B_i + \max_{R \in \mathbf{R}} \sum_{i \in top\ border\ of\ R} B_i +$$
$$\max_{R \in \mathbf{R}} \sum_{i \in right\ border\ of\ R} B_i + \max_{R \in \mathbf{R}} \sum_{i \in bottom\ border\ of\ R} B_i \Big) \qquad (4)$$

Finally, we get a refined bounding box, specifying active region on a magn-image more precisely. The result of the procedure with $\alpha = 4, \beta = 0.05$ is presented in figure figure 2 a,

4.2.3 Neutral line

The neutral line concept is supposed to give several informative features. We define it as follows: *neutral line* is a line separating the regions of strong magnetic fields obtained from segmentation with different polarities.

Neutral line extraction is applied to a refined active region. We use both direct implementation of the definition and the robust algorithm that cuts off small regions and small fragments of the line. The resulting simple neutral line is presented in white in figure 2 b, the robust line is given in black in the same figure.

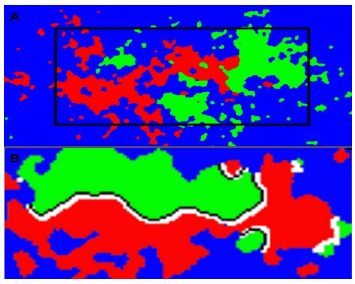

Figure 2: (a) — the localized active region, (b) — standard (white) and robust (black) versions of the neutral line.

4.2.4 Extracted features

Magn-features are calculated using head magnetogram image, their speed of change — using the both images. An active region is unambiguously defined by the solar flare, corresponding to a precedent. An AR is specified by it's bounding box.

Extracted magn-features:
1. sum of magnetogram values in the area corresponding to a positive/negative segment inside bounding box (positive/negative flux),
2. maximum absolute value of the magnetic flux in the bounding box (maximum absolute flux),
3. area of the dilated line with the dilation coefficient equal to 5 for simple and robust algorithms (simple/robust line area),
4. sum of the positive/negative magnetic flux inside the area of the dilated line with the dilation coefficient equal to 5 for simple and robust algorithms (simple/robust line positive/negative flux),
5. speed of change for all of the features above.

5. TESTING SYSTEM AND RESULTS

For testing purposes we use libSVM implementation of Support Vector Machine two-class classifier. After several numerical experiments with different kernel and structural parameters we have found that the best result of SVM with RBF-kernel is worse than with linear one. So, further only linear SVM with the only structural parameter C is used. We implemented an exhaustive search of the optimal parameter value over the set $C_{set} = \{2.25^i \,|\, i \in \{-4, -3, \ldots, 4\}\}$.

Several precedents can have correspondence to one flare. So, we can define max-based decision rule: we obtain class labels for all test precedents in usual way, then we perform postprocessing organized as follows. For every precedent we calculate maximum among class labels of the precedents, which correspond to the same flare as our precedent. Thus, we decrease probability of missing positive precedents.

Every calculated dataset $(X_{M_f,T_f}, Y_{M_f,T_f})$ is divided into three non-intersecting approximately equal parts: $train$ (to learn our classifier), $test$ (to find the most optimal configurations of the structural parameters), and $TEST$ (to get testing results). The union of $train$ and $test$ is denoted as $TRAIN$.

Although negative precedents are much more numerous than positive (strong flare is a rare event), we should learn and optimize our classifier on balanced datasets $train$ and $test$.

Except this, to get rid of the similarity between the precedents in one series of flares, we decided to put the precedents, corresponding to flares belonging to the same flare series, either all in train or all in test.

Although a reasonable amount of features is calculated, we don't exactly know, which of them are really informative for one or the other precedent model configuration. Therefore, we implemented full search over all subsets $\{F \,|\, F \subseteq \mathbf{2}^{F_{set}}, F \neq \emptyset\}$. F_{set} includes the following features: umbra square of the sunspot group and it's speed of change, negative flux and it's speed of change, maximum absolute flux and it's speed of change, robust line negative flux and it's speed of change, speed of change of the simple line negative flux.

To obtain the final results table 1 we run our testing procedure for every configuration $(M_f, T_f) \in \{M_{set} \times T_{set}\}$; partitioning into $TRAIN$ and $TEST$ was fixed; the results were averaged over 5 different partitions of $TRAIN$ into $train$ and $test$.

M_f \ T_f	1.00	1.50	2.00
C5.0	36.0 **26.9**	36.6 **29.3**	37.0 **30.1**
M1.0	31.7 **22.3**	36.2 **22.2**	36.5 **22.2**
M5.0	26.7 **12.3**	29.3 **13.7**	25.7 **13.7**
X1.0	19.9 **9.4**	17.3 **10.2**	19.0 **11.6**

Table 1: Average error rates (%) on $TEST$ dataset: for balanced (first number) and unbalanced (second number) it's versions.

Since real data are unbalanced (strong flares are much less than weak ones), second number (bold) in every cell can be considered as unbiased error rate of our method; a cell is chosen according our forecast needs. Finally, we have 63% to 82% on balanced data (maximum worst-case), 73% to 90% on real data.

In the nearest future we intend to incorporate in our method some additional physically-driven features, include SHO/HMI images support, and build fully-automated web-compliant prediction system.

6. REFERENCES

[1] Zhongxian Shi and Jingxiu Wang, "Delta-sunspots and x-class flares," *Solar Physics*, vol. 149, pp. 105–118, jan 1994.

[2] Patrick S. McIntosh, "The classification of sunspot groups," *Solar Physics*, vol. 125, pp. 251–267, 1990.

[3] T. Colak and R. Qahwaji, "Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares," *Space Weather*, vol. 7, no. 6, jun 2009.

[4] Daren Yu, Xin Huang, Huaning Wang, Yanmei Cui, Qinghua Hu, and Rui Zhou, "Short-term solar flare level prediction using a bayesian network approach," *The Astrophysical Journal*, vol. 710, no. 1, pp. 869, 2010.

[5] David Falconer, Abdulnasser Barghouty, and et. al, "A tool for empirical forecasting of major flares, coronal mass ejections, and solar particle events from a proxy of active-region free magnetic energy," *Space Weather*, vol. 9, no. 4, apr 2011.

[6] S. Zharkov, V. Zharkova, and et. al, "Automated recognition of sunspots on the soho/mdi white light solar images," in *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 3215, pp. 446–452. 2004.

[7] Michael I. Jordan, Zoubin Ghahramani, and et. al, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, pp. 183–233, November 1999.

[8] D. Kropotov, D. Laptev, and et. al, "Variational segmentation algorithms with label frequency constraints," *Pattern Recognit. Image Anal.*, vol. 20, pp. 324–334, September 2010.

[9] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Computer Vision and Pattern Recognition, 2008. IEEE Conference*, 2008, pp. 1–8.

7. ACKNOWLEDGEMENTS

This work is supported by Microsoft Research grant. Authors would like to thank Anatoly Petrukovich, Victor Lempitsky and Pushmeet Kohli for valuable discussions throughout the research.