2022

Building a "parsimonious model" for an effective prediction of solar flare

DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF BRADFORD

FAHIM RAFIQUE, UB: 21023611

Abstract

A solar flare is a sudden explosion of energy caused by the crossing or tangling of magnetic field lines near sunspots. By investigating the association between sunspot properties and flare occurrence, the aim of this paper is to develop an AI prototype system that can predict the correlation between flares (type C, M, or X-class) and find another dataset with less features that can provide better prediction (Colak and Qahwaji, 2009) using the solar flare dataset provided by the UCI Machine Learning Repository.

Introduction

A lot of activity takes place on the surface of the Sun. Electricity charges gases on its surface, creating powerful magnetic field. Gases in the Sun are constantly moving, which tangles, stretches, and twists the magnetic fields. As a result, there is a lot of activity on the Sun's surface, called solar activity. On the surface of the Sun, sunspots appear dark. Their dark appearance is due to the fact that they are cooler than other parts of the solar surface. Near sunspots, magnetic field lines are often tangled, crossed, and reorganized. As a result, a sudden burst of energy is generated, known as a solar flare. Solar flares emit a lot of radiation into space. The radiation released by a solar flare can interfere with our radio communications here on Earth if it is very intense(spaceplace,2021). In addition, solar flares can cause severe damage to our infrastructure, degrading the Global Positioning System (GPS), interfering with power grids, and causing problems with communications satellites. These solar storms can also cause GPS data to be inaccurate. A commercial plane relies on GPS to take off, navigate, and land. Space weather currents produced in the ionosphere could generate huge currents within power grids, resulting in the damage of the transformers that form the backbone of these systems. Finally, ionizing particle radiation from flares and CMEs can cause damage to or even result in the loss of communications satellites. This was the case with Galaxy 15 in April 2010. To prepare safely for damaging space weather, operations teams need a precise prediction of solar flares to be able to perform their jobs: power grid operators need to know when ionospheric currents are expected, satellite operators need to know when to shut down equipment; pilots need to know when to divert transpolar flights to lower latitudes; astronauts need to know when to seek cover in shielded areas.

Background

Artificial Intelligence is a branch of Computer Science that uses models and methods to solve problems and make predictions using massive amounts of data. Artificial intelligence consists of three main parts:

an agent, an environment, and a goal. In addition to giving the environment actions, the agent must also receive input from the environment. Additionally, the agent must have a goal that it is trying to achieve and, in order to make its goal flexible and not limiting, the agent must receive another input, reward. The agent's goal will then be to maximize its reward (Legg &Hutter, 2018).

Protecting crucial space and ground infrastructure from extreme solar radiation requires the development of AI systems that can predict such events in advance.

The dataset we are given uses 3 component McIntosh classification which is based on the general form 'Zpc', where 'Z' is the modified Zurich Class, 'p' describes the penumbra of the principal spot, and 'c' describes the distribution of spots in the interior of the group(McIntosh,1990).

Z-values: (Modified Zurich Sunspot Classification)

A - A small single unipolar sunspot. Representing either the formative or final stage of evolution.

B - Bipolar sunspot group with no penumbra on any of the spots.

C - A bipolar sunspot group. One sunspot must have penumbra.

D - A bipolar sunspot group with penumbra on both ends of the group. Longitudinal extent does not exceed 10 deg.

E - A bipolar sunspot group with penumbra on both ends. Longitudinal extent exceeds 10 deg. but not 15 deg.

F - An elongated bipolar sunspot group with penumbra on both ends. Longitudinal extent of penumbra exceeds 15 deg.

H - A unipolar sunspot group with penumbra.

p-values:

X - no penumbra (group class is A or B)

R - rudimentary penumbra partially surrounds the largest spot.

This penumbra is incomplete, granular rather than filamentary, brighter than mature penumbra, and extends as little as 3 arcsec from the spot umbra. Rudimentary penumbra may be either in a stage of formation or dissolution.

S - small, symmetric (like Zurich class J). Largest spot has matured, dark, filamentary penumbra of circular or elliptical shape with little irregularity to the border. The north-south diameter across the penumbra is less or equal than 2.5 degrees.

A - small, asymmetric. Penumbra of the largest spot is irregular in outline and the multiple umbrae within it are separated. The north-south diameter across the penumbra is less or equal than 2.5 degrees.

H - large, symmetric (like Zurich class H). Same structure as type 'S', but north-south diameter of penumbra is more than 2.5 degrees. Area, therefore, must be larger or equal than 250 millionths solar hemisphere.

K - large, asymmetric. Same structure as type 'A', but north-south diameter of penumbra is more than 2.5 degrees. Area, therefore, must be larger or equal than 250 millionths solar hemisphere.

c-values:

X - undefined for unipolar groups (class A and H)

O - open. Few, if any, spots between leader and follower. Interior spots of very small size. Class E and F groups of 'open' category are equivalent to Zurich class G.

I - intermediate. Numerous spots lie between the leading and following portions of the group, but none of them possesses mature penumbra.

C - compact. The area between the leading and the following ends of the spot group is populated with many strong spots, with at least one interior spot possessing mature penumbra. The extreme case of

compact distribution has the entire spot group enveloped in one continuous penumbral area.

Predicting solar flares is extremely complex and tedious process and thus requires lot of resource and experts. So, AI could help to process the data and take the burden away from using vital resources.

Method

In this section, we are going to provide the detailed steps by which we can find dataset with less features that can provide better prediction for the solar flares. Also, we will try to find the correlation between the C, M and X class flares.

We are going to do a Regression analysis to find the significant variables that we can use to predict the C, M or X class flares.

As a mathematical process, regression analysis involves estimating relations between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). Linear regression is the most common form of regression analysis, which allows researchers to find the line (or a more complex linear combination) that most closely matches the data according to some mathematical criterion.

Since the independent variables such as, largest spot size, spot distribution, activity, evolution, previous 24-hour flare activity, historically complex, area, did region become historically complex on this pass across the sun's disk, area and area of the largest spot are nominal/categorical variable and the predictor/dependent variables C, M, X class flares are continuous variables; linear regression model will not work in this case. So, we are going to use Ordinal Regression model to find the significant variables.

However, Ordinal regression does not work for ‘CHAR’/string variables. Therefore, we are going to use pseudo-code for the largest spot size and spot distribution. So, we renamed A, H, K, R, S, X spot size as 1,2,3,4,5,6 respectively. Moreover, we do the same for the spot distribution and rename C, I, O, X as 1,2,3,4 respectively.

To find the relationship between the C-class flares, M-class flares and the X-class flares, we will do a Pearson correlation coefficient(2-tailed) between them. A correlation coefficient measures the strength of the association between the relative movements of two variables. The range of values is between -1.0 and 1.0. If the calculated value is greater than 1.0 or less than -1.0, an error occurred in the calculation.

When the correlation is -1.0, there is a true negative correlation, and when it is 1.0, there is a true positive correlation. A correlation of 0.0 means that there is no correlation between the variables.

A Pearson product-moment correlation coefficient, also known as r, R or Pearson's r, measures the strength and direction of a linear relationship between two variables. It is defined as the product of the covariances of the variables divided by the standard deviations of the variables. This is the best-known and most used type of correlation coefficient.

Analysis

Data Processing Steps Taken:

1. Download dataset from http://archive.ics.uci.edu/ml/datasets/solar+flare
2. Extract dataset to the folder using: tar xvf solar-data.tar.z
3. Providing pseudo-code for largest spot size A, H, K, R, S, X as 1,2,3,4,5,6 and spot distribution C,I, O, X as 1,2,3,4.

We used C-class flares, M-class flares and, X-class flares as the outcome variables and largest spot size, spot distribution, activity, evolution, previous 24-hour flare activity, historically complex, area, did region become historically complex on this pass across the sun's disk, area and area of the largest spot as independent/predictor variable for the Ordinal Regression.

Our regression model gives the following results:

We can see from the regression,

1. there are no significant variables to predict the X-class flares p>**0.05**.
2. For the M-class flares, spot distribution(C,I) is significant because their significance level p<**0.05.**
3. For the C-class flares, largest spot size(A,H,K,S), spot distribution(C,I,O), activity(1) is significant variable(p<**0.05**).

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[X_class_flares = 0]	17.594	187.841	.009	1	.925	-350.569	385.756
	[X_class_flares = 1]	19.372	187.844	.011	1	.918	-348.795	387.538
Location	[largest_spot_size=1]	6.134	140.417	.002	1	.965	-269.079	281.346
	[largest_spot_size=2]	-2.593	389.716	.000	1	.995	-766.422	761.236
	[largest_spot_size=3]	6.375	140.422	.002	1	.964	-268.847	281.597
	[largest_spot_size=4]	-.829	181.903	.000	1	.996	-357.353	355.695
	[largest_spot_size=5]	-.908	165.466	.000	1	.996	-325.215	323.399
	[largest_spot_size=6]	0[a]	.	.	0	.	.	.
	[spot_distribution=1]	8.388	124.773	.005	1	.946	-236.161	252.938
	[spot_distribution=2]	6.770	124.768	.003	1	.957	-237.771	251.311
	[spot_distribution=3]	-.417	148.963	.000	1	.998	-292.378	291.544
	[spot_distribution=4]	0[a]	.	.	0	.	.	.
	[activity=1]	-.567	1.328	.182	1	.669	-3.171	2.037
	[activity=2]	0[a]	.	.	0	.	.	.
	[evolution_2=1]	-7.652	193.861	.002	1	.969	-387.611	372.308
	[evolution_2=2]	-.347	1.132	.094	1	.759	-2.565	1.872
	[evolution_2=3]	0[a]	.	.	0	.	.	.
	[previous_24_hour_flare_activity_code_2=1]	-.018	1.237	.000	1	.988	-2.441	2.406
	[previous_24_hour_flare_activity_code_2=2]	-11.108	277.024	.002	1	.968	-554.066	531.850
	[previous_24_hour_flare_activity_code_2=3]	0[a]	.	.	0	.	.	.
	[historically_complex=1]	-6.699	69.212	.009	1	.923	-142.352	128.955
	[historically_complex=2]	0[a]	.	.	0	.	.	.
	[did_region_become_historically_complex_on_this_pass_across_the_s=1]	.542	197.335	.000	1	.998	-386.228	387.312
	[did_region_become_historically_complex_on_this_pass_across_the_s=2]	0[a]	.	.	0	.	.	.
	[area_2=1]	1.988	1.721	1.334	1	.248	-1.386	5.361
	[area_2=2]	0[a]	.	.	0	.	.	.
	[area_of_the_largest_spot=1]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Threshold	[M_class_flares = 0]	7.471	1.597	21.876	1	.000	4.340	10.602
	[M_class_flares = 1]	9.216	1.636	31.723	1	.000	6.009	12.423
	[M_class_flares = 2]	9.792	1.669	34.434	1	.000	6.521	13.063
	[M_class_flares = 3]	10.498	1.741	36.344	1	.000	7.085	13.911
	[M_class_flares = 4]	11.201	1.879	35.548	1	.000	7.519	14.883
Location	[largest_spot_size=1]	1.998	1.105	3.272	1	.070	-.167	4.164
	[largest_spot_size=2]	1.582	1.489	1.129	1	.288	-1.336	4.500
	[largest_spot_size=3]	1.876	1.325	2.006	1	.157	-.720	4.472
	[largest_spot_size=4]	.813	1.169	.483	1	.487	-1.479	3.104
	[largest_spot_size=5]	1.725	1.082	2.543	1	.111	-.395	3.845
	[largest_spot_size=6]	0[a]	.	.	0	.	.	.
	[spot_distribution=1]	3.212	1.252	6.577	1	.010	.757	5.666
	[spot_distribution=2]	2.237	1.092	4.193	1	.041	.096	4.378
	[spot_distribution=3]	1.679	1.065	2.487	1	.115	-.408	3.766
	[spot_distribution=4]	0[a]	.	.	0	.	.	.
	[activity=1]	.485	.499	.945	1	.331	-.492	1.462
	[activity=2]	0[a]	.	.	0	.	.	.
	[evolution_2=1]	.022	1.096	.000	1	.984	-2.125	2.169
	[evolution_2=2]	.818	.409	4.003	1	.045	.017	1.619
	[evolution_2=3]	0[a]	.	.	0	.	.	.
	[previous_24_hour_flare_activity_code_2=1]	.352	.761	.214	1	.644	-1.139	1.843
	[previous_24_hour_flare_activity_code_2=2]	.536	1.195	.201	1	.654	-1.806	2.879
	[previous_24_hour_flare_activity_code_2=3]	0[a]	.	.	0	.	.	.
	[historically_complex=1]	-.292	.454	.413	1	.520	-1.182	.598
	[historically_complex=2]	0[a]	.	.	0	.	.	.
	[did_region_become_historically_complex_on_this_pass_across_the_s=1]	.201	1.427	.020	1	.888	-2.596	2.998
	[did_region_become_historically_complex_on_this_pass_across_the_s=2]	0[a]	.	.	0	.	.	.
	[area_2=1]	.884	.843	1.099	1	.294	-.769	2.537
	[area_2=2]	0[a]	.	.	0	.	.	.
	[area_of_the_largest_spot=1]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Threshold	[C_class_flares = 0]	3.351	.558	36.011	1	.000	2.256	4.445
	[C_class_flares = 1]	4.587	.568	65.165	1	.000	3.473	5.700
	[C_class_flares = 2]	5.324	.580	84.303	1	.000	4.188	6.461
	[C_class_flares = 3]	6.161	.607	103.000	1	.000	4.971	7.351
	[C_class_flares = 4]	6.939	.659	110.733	1	.000	5.646	8.231
	[C_class_flares = 5]	7.641	.748	104.370	1	.000	6.175	9.107
	[C_class_flares = 6]	9.038	1.145	62.348	1	.000	6.794	11.281
Location	[largest_spot_size=1]	1.415	.396	12.742	1	.000	.638	2.192
	[largest_spot_size=2]	1.372	.603	5.180	1	.023	.191	2.554
	[largest_spot_size=3]	2.331	.542	18.466	1	.000	1.268	3.394
	[largest_spot_size=4]	.621	.404	2.364	1	.124	-.170	1.411
	[largest_spot_size=5]	1.149	.379	9.209	1	.002	.407	1.891
	[largest_spot_size=6]	0[a]	.	.	0	.	.	.
	[spot_distribution=1]	1.685	.566	8.876	1	.003	.576	2.793
	[spot_distribution=2]	1.858	.380	23.932	1	.000	1.114	2.603
	[spot_distribution=3]	1.166	.351	11.024	1	.001	.478	1.854
	[spot_distribution=4]	0[a]	.	.	0	.	.	.
	[activity=1]	-.592	.229	6.658	1	.010	-1.041	-.142
	[activity=2]	0[a]	.	.	0	.	.	.
	[evolution_2=1]	-.471	.485	.941	1	.332	-1.422	.481
	[evolution_2=2]	.086	.186	.213	1	.644	-.279	.450
	[evolution_2=3]	0[a]	.	.	0	.	.	.
	[previous_24_hour_flare_activity_code_2=1]	.080	.445	.032	1	.857	-.793	.953
	[previous_24_hour_flare_activity_code_2=2]	-.014	.645	.000	1	.983	-1.277	1.249
	[previous_24_hour_flare_activity_code_2=3]	0[a]	.	.	0	.	.	.
	[historically_complex=1]	.026	.214	.014	1	.905	-.395	.446
	[historically_complex=2]	0[a]	.	.	0	.	.	.
	[did_region_become_historically_complex_on_this_pass_across_the_s=1]	-.185	.504	.135	1	.713	-1.172	.802
	[did_region_become_historically_complex_on_this_pass_across_the_s=2]	0[a]	.	.	0	.	.	.
	[area_2=1]	-.269	.543	.246	1	.620	-1.333	.795
	[area_2=2]	0[a]	.	.	0	.	.	.
	[area_of_the_largest_spot=1]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Since we could not find any significant variable to predict X-class flare, we are going to find the correlation of X-class with the other two kind of flares to predict measure their association. Therefore, we will do a Pearson's correlation coefficient between them.

From the correlation matrix, it is evident that

1. C & M class flares are positively correlated, and correlation is significant.
2. C & X class flares are positively correlated but the correlation is not significant.
3. M & X class flares positively correlated, and correlation is significant.

Correlations

		C_class_flares	M_class_flares	X_class_flares
C_class_flares	Pearson Correlation	1	.148**	.029
	Sig. (2-tailed)		.000	.352
	N	1066	1066	1066
M_class_flares	Pearson Correlation	.148**	1	.420**
	Sig. (2-tailed)	.000		.000
	N	1066	1066	1066
X_class_flares	Pearson Correlation	.029	.420**	1
	Sig. (2-tailed)	.352	.000	
	N	1066	1066	1066

**. Correlation is significant at the 0.01 level (2-tailed).

****all the calculation is done on IBM-SPSS (version 25.0)**

Conclusion

In conclusion, from the Ordinal regression model we can say that area and spot distribution is a good predictor to predict M-class flares. Also, largest spot size, spot distribution, activity are good predictors for C-class flares. So, we can predict these two solar flares from smaller number of predictors. Since we could not find any good predictor for X-class flares, so we did a Pearson's correlation test to see if there any correlation between the flares and found out that M & X-class flares are positively correlated, and correlation is significant.

Further investigation can be done from the dataset in the future such as: prediction of no flare happening, prediction of significant flares, find a pattern of sunspots to predict the significant flares.

Reference

Legg, S. and Hutter, M., 2007. Universal intelligence: A definition of machine intelligence. *Minds and machines*, *17*(4), pp.391-444.

Spaceplace.nasa.gov. 2022. *Sunspots and Solar Flares | NASA Space Place – NASA Science for Kids*. [online] Available at: <https://spaceplace.nasa.gov/solar-activity/en/> [Accessed 6 January 2022].

McIntosh, P., 1990. The classification of sunspot groups. *Solar Physics*, 125(2), pp.251-267.

Colak, T. and Qahwaji, R., 2009. Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares. *Space Weather*, 7(6), p.n/a-n/a.