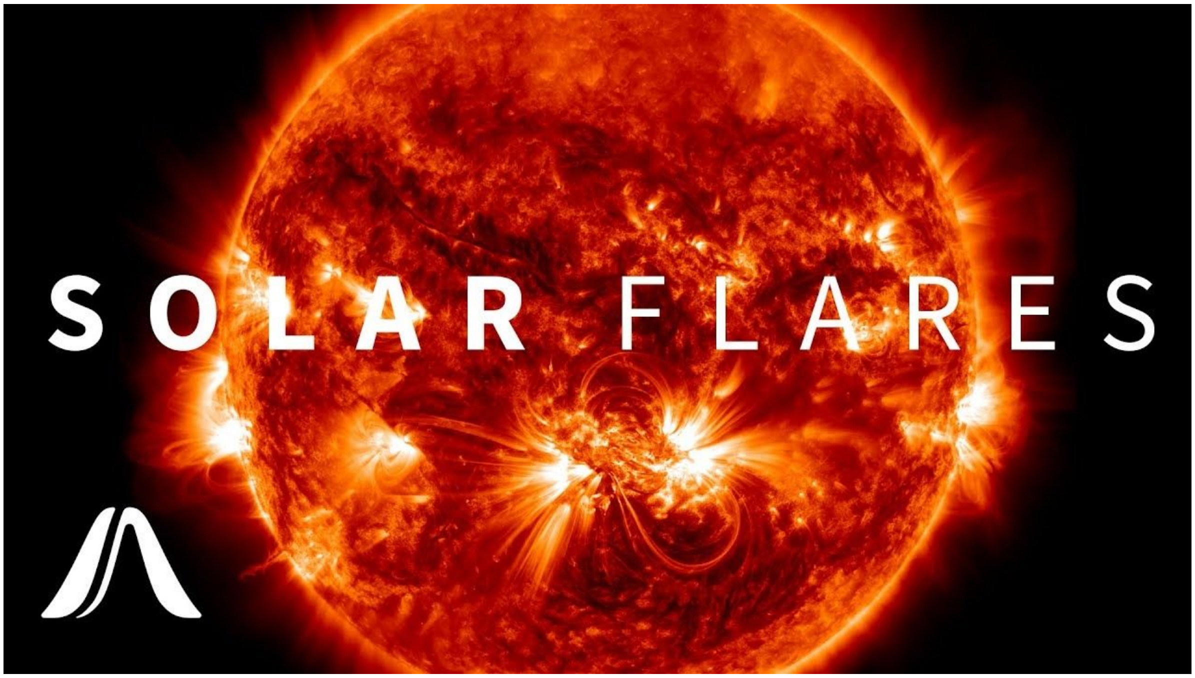


SOLAR FLARE PREDICTION

Using Machine learning & neural network

Aditya(17122003) | Lab Based Project (PHN-300) | Aman Kumar(17122005)

CONTENTS :

Introduction to Solar Flares

Solar flare are sudden explosions of energy generally caused by movements in the magnetic field lines like crossing, tangling or reorganization of these lines. Solar flares generally occur near sunspots. Solar flares can or cannot have mass ejections associated with them. These mass ejections are called **Coronal Mass Ejection**. CME or Coronal Mass Ejection is the ejection of plasma along with the accompanying magnetic field. The plasma ejected generally contains protons and electrons along with various waves.

The classification system for solar flares uses the letters A, B, C, M or X, according to the peak flux in watts per square meter (W/m^2) of X-rays with wavelengths 100-800 picometers.

Classification	**A**	**B**	**C**	**M**	**X**
Peak Flux Range(W/m^2)	$<10^{-7}$	10^{-7} - 10^{-6}	10^{-6} - 10^{-5}	10^{-5} - 10^{-4}	$>10^{-4}$

An earlier flare classification was based on H- α spectral observations. It uses both intensity and emitting surface. The classification in intensity is qualitative, referring to the flares as: faint(f), normal(n) or brilliant(b). The emitting surface is measured in terms of millionths of hemisphere. The total hemisphere area

$$A_H = 15.5 \times 10^{12}\ km^2$$

Classification	**S**	**1**	**2**	**3**	**4**
Corrected Area (Millionths of hemisphere)	<100	100-250	250-600	600-1200	>1200

Smaller flares (C-class and some M-Class flares) cause Aurora at poles and are important for astronomical research. Whereas Solar flares on a large scale (X-class and most of the M-class flares) can affect our communication devices and cause blackouts of HF radio communication. So it is very necessary to know about these events in advance to take some action and prevent these Radio blackouts and damage to our communication devices.

AIM

The aim of our project is "**Prediction of solar flares**" using existing data of previous observations as well as classification of predicted flare.

DATASET

The original source of data in our data set is from the **National Ocean and Atmospheric Administration(NOAA)** under the U.S. Department of Commerce. Data originates from the National Geophysical Data Center(NGDC) which is now called National Center of Environmental Information(NCEI).

Link to the original source is: ftp://ftp.swpc.noaa.gov/pub/warehouse

We have used the data from **June 1997 to May 2020** having daily observations. Data available earlier from this timeline exists just as records of flare events and no further Tabulated data is available so it cannot be used to predict solar flare events. The data obtained from the website is in the form pdfs which has various other information and reports along with our required data. Required data has been filtered out from the pdf files and then clustered together initially.

Firstly, we have to specify which fields are the required fields of data for prediction. This can be done based on the definition and existing known correlating factors with a solar flare. After analyzing the data source, we get the fields which are namely:

1. **Date**: Contains the date month and year of the flare occurrence. It is required so that we can predict the occurrence of solar flare on a specific date.
2. **Radio Flux**: This column records the change in average Radio flux (λ=10.7 cm) coming from the sun on a specific date.
3. **Sunspot Number**.: It records the number of sunspots occurring on a single day.
4. **Sunspot Area**: Average area covered by sun spots every day (in hemispheric relative spot number).
5. **X-ray Background Flux:** It contains the type (A, B, C – here A is least high value and C is most) and scale value of change in the background x-ray flux.
6. **Proton Fluence:** It contains the mass ejection data for protons. It has three columns for three different energy level ranges namely :

   >1 Mev, >10 Mev, >100 Mev

   (These also correspond to different mass ranges).

7. **Electron Fluence:** It contains the mass ejection data for electrons. It has three columns for three energy ranges namely: > 0.6 MeV, > 2 MeV, > 4 MeV but generally measured values are in > 2 MeV range hence the other energy ranges have been removed for the convenience and accuracy of modelling.
8. **K-Indices:** It denotes the disturbances in horizontal magnetic field lines of earth's atmosphere (Scaled 0-9 where after 5 denotes geomagnetic storms). We have data for different altitudes.
9. **Classifications**: It contains the two major classes of flares namely X-ray flux and Optical as well as their sub-classes. Namely C, M, X for X-ray flux and S,1,2,3,4 for Optical ones.

This data is then converted into an excel file. The data set so formed also contains various missing values and non-segregated variables which have to be further processed.

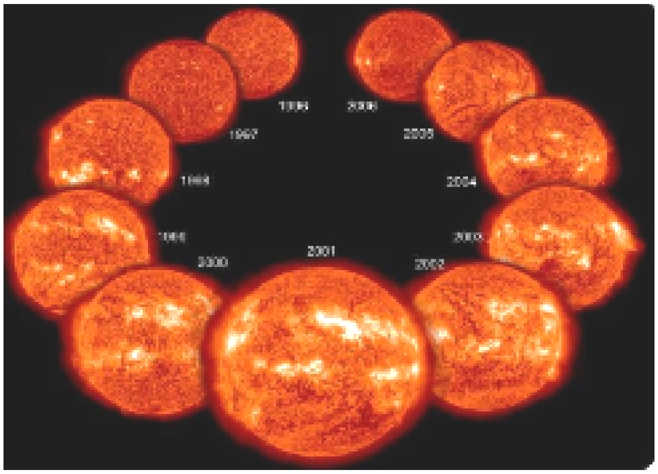
Dataset Visualisation

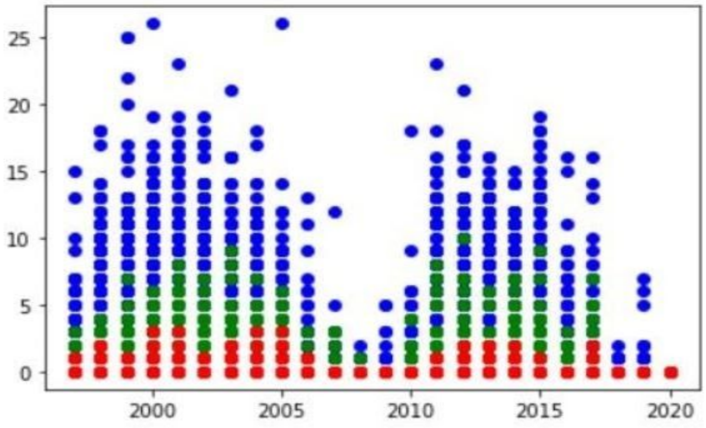
The Sun's magnetic field goes through a cycle, called the **solar cycle**. Every 11 years or so, the Sun's magnetic field completely flips. This affects the activity on the surface of the Sun, such as sunspots and the solar flares. During this solar cycle the activity of the sun from a minima to a maxima and then back to minima. So the occurrence of solar flares is a kind of periodic phenomenon. This can also be seen through our dataset. Our dataset is spread over a span of 23 years (i.e. two solar cycles).

We have plotted the occurrence of solar flares (Right figure).

x-axis: Time || y-axis: Number of flares occurring per day

Blue: C-class flares || **Green**: M-class flares || **Red**: X-class flares





PRE-PROCESSING

The tabulated data obtained in the form of excel file has various missing values and non-segregated columns. We process them, firstly we segregate the values of k indices and give them independent columns according to altitudes and types. We also do the same for Flare Classes. Then comes the most crucial step which is filling the missing values. Astronomical data has a lot of inherited missing values in it due to telescopic or mechanical or atmospheric problems i.e. a lot of data is and can be missing. Here in the dataset we have a lot of columns so some data is missing which cannot be filled using normal averaging techniques like athematic mean, median values etc. Hence we have to fill out the missing values using a different method.

We consider each data point as points in a vector space. So points which are closer have similar values in their columns. We fill the missing values on a particular day as:

To fill the missing values we consider only those data points whose values were not missing. We take a total of 360 data points to find the datapoint which is nearest to it in our vector space. These data points are selected as : (i) 180 before and 180 after our data point (ii) 360 after or before our datapoint if the data points are not available before or after (respectively) our data point. We then get the normalized difference of column values of our datapoint and the selected datapoint and then calculated the square root of the sum of squared values of this normalized difference. The datapoint which has the least value of the calculated distance and fills the missing value of our datapoint from it.

We also made different modifications in the dataset according to model requirements.

Most of the time of our project is consumed in preparing and preprocessing the dataset. It took a lot of efforts to make this dataset which is well structured and cleaned.

MODELS AND ARCHITECTURES

We have used different models to analyze and predict solar flares. We have used several regression models for establishing relations between flare classes and other variables. As it is a classification problem i.e. we classify which type and class of solar flare occurrence a particular event is based on other variables hence we use classification models. We also tried to use Artificial Neural Networks (ANNs) and Recurrent Neural Networks(RNNs) on a trial and error basis to get better results of which we are only including RNNs results in this report as others were unsatisfactory.

We also tried to find a relationship between occurrence of one class flares to other class flares. So we built several sub-models using different data:

Flare Class	Number of models	Data Used
C-class	1	
M-class	2	(i) Without using C-class flares data (ii) Using C-class flares data
X-class	3	(i) Without using C & M class flares data (ii) Using C-class flares data (iii) Using C & M class flares data

The models we build are:

a) K Nearest Neighbours (KNN): We used KNN for classification type trials and tested it with different parameters. We have built two KNN classifiers (i)For classifying each flare class individually (ii)For classifying all flares at once. We used different values for **n_neighbours** (3,5,10) and **train-to-test ratio** (8:2, 7:3, 6:4). Other parameters used were :

```
KNeighborsClassifier(algorithm='auto', leaf_size=30,
metric='minkowski',metric_params=None, n_jobs=None, n_neighbors=__,
p=2,weights='uniform')
```

These parameters were either default or chosen after various hit and trials

b) Support Vector Regression (SVR): It is a method in Support Vector Machines which as the name suggests runs Regression models. We tried SVR with three **kernels** namely: (i)**Linear** (ii)**Polynomial** and (iii)**Radial Basis Function** (**RBF**). The other parameter values were given their default values.

```
SVR(C=1.0, cache_size=200, coef0=0.0, degree=3, epsilon=0.1,tol=0.001,
verbose=False, gamma='scale',kernel=’___', max_iter=-1, shrinking=True)
```

c) Random Forest Classifier: We used Random forest for classification type trials and tested it with different parameters. Parameters used were :

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None,
max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0,
n_estimators=100, n_jobs=None, oob_score=False, random_state=0, verbose=0,
warm_start=False)
```

These parameters were either default or chosen after various hit and trials.

d) **Naive Bayes:** We used Naive Bayes for classification type trials and tested it with different parameters. Parameters used were :

```
GaussianNB(priors=None, var_smoothing=1e-09)
```

These parameters were either default or chosen after various hit and trials.

e) **RNN (Using LSTM cells):** We used LSTM cells to build a Recurrent Neural Network. To predict solar flares on a particular day we used its previous 30 days data. We have created a data structure to contain the data of 30 days. After doing a lot of training on different values of these parameters we find the model gives optimum results for these following values:

```
# Parameters

n = 30      # Number of days

t = 6704    # Training set size (80% of total dataset)

m = 5       # No. of hidden LSTM layers(except first and last)

u = 30      # Number of nodes in each hidden layer

d = 0.2     # Dropout fraction

b = 32      # Batch size

e = 10      # Number of epochs
```

The model contains 5 hidden layers and total 7 layers including the first input layer and last output layer. The epochs have been kept low because:
(i) Accuracy ceases to increase and oscillates around a particular value after e=10.
(ii) Large numbers of epochs took enormous amounts of time to complete.

RESULTS

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \textit{Precision} = \frac{TP}{TP + FP}$$

$$\textit{Recall} = \frac{TP}{TP + FN} \qquad \textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

TN : True Negative | FN : False Negative | TP : True Positive | FP : False Negative

Models	Sub Model	Flare Class	Accuracy	Precision	Recall	f1-score	

KNN	For Individual Flares	for C-class
		for M-class
		for X-class
			0.82
			0.85
			0.98
				0.86
				0.48
				0.75
					0.78
					0.26
					0.05
						0.82
						0.34
						0.09
							n=5 | train:test = 6:4
							n=3 | train:test = 6:4 |Without C
							n=5 | train:test = 6:4 |With C&M.

	For all X-ray class flares	for C-class
		for M-class
		for X-class
			0.67	0.59
				0.45
				0.67
					0.60
					0.16
					0.07
						0.59
						0.24
						0.13
							n=5 | train:test = 8:2

SVR	RBF kernel	for C-class
		for M-class
		for X-class
			0.86
			0.86
			0.98
				0.87
				0.76
				0.50
					0.85
					0.12
					0.04
						0.86
						0.20
						0.07
							Without using C-class flare data.
							Using C & M class flare data.

	Polynomial kernel	for C-class
		for M-class
		for X-class
			0.84
			0.86
			0.98
				0.84
				0.65
				0.25
					0.84
					0.11
					0.07
						0.84
						0.18
						0.11
							Without using C-class flare data.
							Without using C & M data.

	Linear kernel	for C-class
		for M-class
		for X-class
			0.85
			0.86
			0.98
				0.84
				0.76
				0.00
					0.88
					0.06
					0.00
						0.86
						0.12
						0.00
							Without using C-class flare data.
							In all Scenarios.

Naive Bayes	-----	for C-class
		for M-class
		for X-class
			0.76
			0.81
			0.88
				0.87
				0.37
				0.08
					0.62
					0.41
					0.55
						0.72
						0.39
						0.15
							Using C-class flare data.
							Using C & M data.

Random forest classifier	-----	for C-class
		for M-class
		for X-class
			0.84
			0.86
			0.98
				0.93
				0.68
				1.00
					0.73
					0.11
					0.04
						0.82
						0.18
						0.07
							Without using C-class flare data.
							Using C-class flares data.

RNN	**LSTM cell**	**for C-class**
		for M-class
		for X-class
			0.73
			0.99
			1.00
				0.36
				1.00
				1.00
					0.81
					0.47
					0.67
						0.50
						0.64
						0.80	

LINK :

Link : https://drive.google.com/drive/folders/15z4tXB8jf1PhQOBQBlafESppUrYXU2Hv?usp=sharing

This Link contains all the dataset files, programs used in creating and preprocessing dataset and Models.