AN INVESTIGATION INTO APPLICATIONS OF MACHINE LEARNING ALGORITHMS ON SOLAR FLARE DATA AND DISTANCE PREDICTION

Isaac Abraham
INSPIRIT AI 1:1 X Independent Researcher Affiliation
Singapore

ABSTRACT

A solar flare is an intense localized eruption of electromagnetic radiation in the sun's atmosphere. A solar flare is typically accompanied with a coronal mass ejection (CME), in which a highly magnetic plasma is released from the sun's corona into the heliosphere. CMEs are capable of reaching earth and colliding with the earth's magnetic field, causing dangerous geomagnetic storms that can start fires and cause power outages. It is imperative that flares are studied so that actions can be taken to mitigate the effect of geomagnetic storms. Our data was sourced from the RHESSI telescope which consisted of almost 100,000 entries[1]. It was shown in Fletcher's paper that there exists a correlation between radial distance and the energy of a flare. We began with classifying solar flares based on their radial distances and then applying machine learning models like logistic regression, KNN classification, Decision Trees and MLP Classification and closely examining the ones that gave the highest accuracy. Our different machine learning models showed that there existed a pattern in the phenomena of solar flares and our AI could be used to predict them accurately. Most notably, the decision tree model had an accuracy of 99.84%.

INTRODUCTION

A solar flare is an intense localized eruption of electromagnetic radiation in the sun's atmosphere. A solar flare is typically accompanied with a coronal mass ejection (CME), which is a release of highly magnetic plasma from the sun's corona into the heliosphere. CMEs are capable of reaching earth and colliding with the earth's magnetic field, causing dangerous geomagnetic storms that can start fires and cause power outages. Solar Flares have resulted in catastrophic events during multiple occasions, one such incident was the Carrington event in 1859, which caused telegraph networks to fail and started fires across the world[2]. A solar flare in our digitalized world could cause even more damage. The nature of CMEs is that they are very difficult to predict, particularly the ones that collide with the earth, a machine learning algorithm could be used to better classify and predict new flares based on current flare data. The dataset that I am using is from the RHESSI (orbiting -extra earth) telescope, which records the solar flares' position, energy, duration and radial distance. The radial distance is the most important factor to consider since CMEs are far more likely to occur on the surface of the sun's corona[3]. The machine learning algorithm is supervised, as the models are trained and then tested using our data.

BACKGROUND

The current method that NASA uses for flare prediction involves studying various solar cycles that range from 11 days to 80 years. However, there are far too many factors to consider using this method of prediction and the forecasts are often wrong. Although, this method is far more logical that giving a machine a set of raw data by using human knowledge to recognize patterns. Another machine learning model is known as Deep Flare Net, which uses image neural networks to extract physical features and then train it using a machine learning model. Although this model is much more accurate than our approach, it requires much higher computational power and even more data to train. The main paper that my research was based on consisted of an in-depth description of flare features and used a unique approach known as the 'multiwavelength

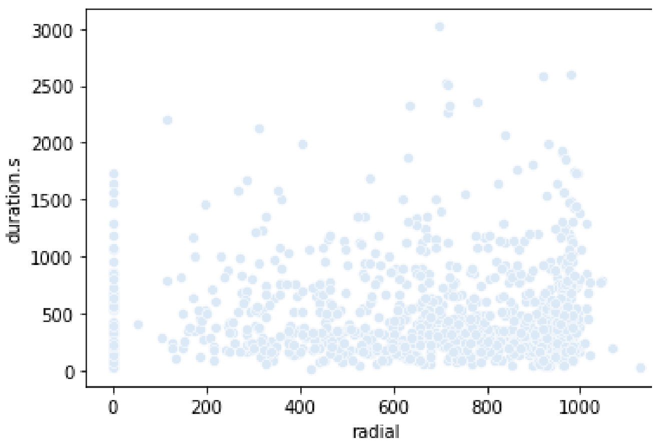
[1] Samaha, Kheirallah. "Solar Flares from RHESSI Mission." *Kaggle*, 7 Apr. 2021, www.kaggle.com/datasets/khsamaha/solar-flares-rhessi.

[2] https://www.space.com/the-carrington-event#:~:text=The%20Carrington%20Event%20was%20a,on%20the%20solar%20disk%20grew.

[3] Dobrijevic, Daisy. "Coronal Mass Ejections: What Are They and How Do They Form?" *Space.com*, Space, 26 May 2022, www.space.com/coronal-mass-ejections-cme.

approach', which measured high energy particles emitted promptly and with delays[4]. This approach could also be useful for determining the composition of solar flares, for example, solar flares are known to consist of a large amount of neutrinos, further insights into this could provide us more data on the behavior of neutrinos and antineutrinos.

[4] Fletcher, L., et al. "An Observational Overview of Solar Flares - Space Science Reviews." *SpringerLink*, Springer Netherlands, 11 Aug. 2011, link.springer.com/article/10.1007/s11214-010-9701-8.



Dataset

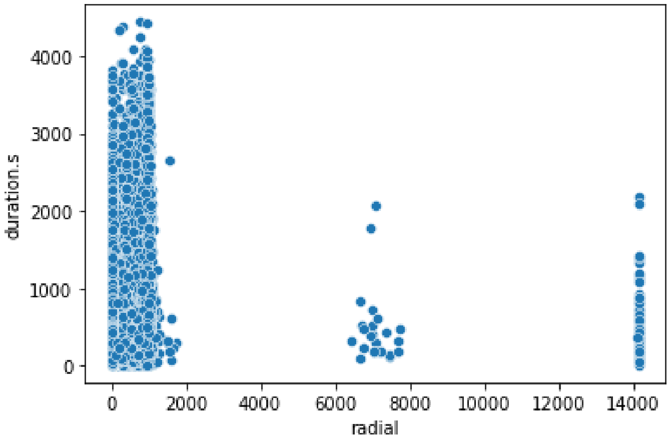
The first step was to format all features of my data into numerical values. I did this by firstly dropping values that I knew would have no correlation with the data, such as the flare index number, date, time and various tags attached to each flare. Additionally, the dataset contained almost 100,000 values.

The next step was to create several radial classification functions as the radial feature only returned a range, non-floating point value. The way I did this was by splitting the range function along the hyphen and storing the mean, lower bound and upper bound into new columns.We applied this transformation as we found that predicting these classifications was both more accurate and more valuable as clear banding of ranges appeared in the dataset. (Maybe show that figure) , since CMEs have a direct correlation with radial distance as they are more likely to form on the sun's corona. The range given in the dataset is due to the movement of these suns in the galaxies we are observing and therefore is periodic. This is why we are able to give a range that is stable in both the upper and lower bound.

I obtained the classification functions by performing a raw dataset analysis from various plots and obtaining the mean and standard deviation of the radius. This allowed me to classify the radial distance d in astronomical units (AU) , where small is 0< d < 400, medium is 400 < d < 700 and large is 700 < d.

Models

I first began by creating multiple scatter plots of variables against each other, such as radial vs duration, radial vs energy and radial vs counts. It showed some correlation, but it was obvious a standard linear regression would not work, as shown in the graph below, which illustrates that several of the variables are not proportional to each other.

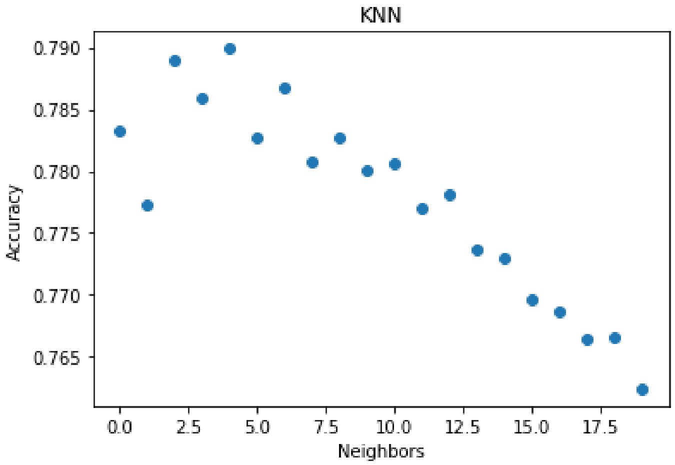


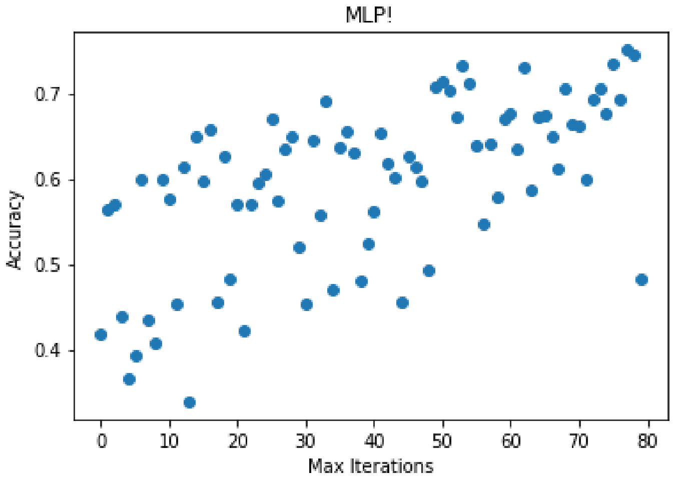
In order to set up my machine learning model, I created a y vector as the radial classification column and the X matrix as all other values. I made sure to remove all the values related to the radial classification such as the lower bound, upper bound and range, including these would have allowed the model to achieve a misleading result.

I then experimented with several different machine learning models, including logistic regression, decision tree, KNN and MLP Classification. These models are ideal for grouping problems where we attempt to create a class label for unlabeled data.

A logistic regression is a model that attempts to classify flares by probability given a set of independent variables. A decision tree model attempts to predict the classification of a new value by creating a tree-like model of decisions to classify the new value. KNN stands for K-nearest neighbors and attempts to classify a new value by examining existing values with similar data. MLP Classification stands for multilayer perceptron classification which attempts to classify a data point by using a neural network, which attempts to mimic the human brain by weighing variables against each other to classify a new datapoint.

Finally, I then attempted to perform hyperparameter optimization on these models by creating a for loop for each of the models and printing the accuracy out for each hyperparameter. After the loop was completed, I then created a scatterplot of hyperparameter vs accuracy for each of these models. For example, For the KNN model, I used a for loop that increased the number of neighbors by one for each iteration.





Results and Discussion

The results showed that the machine learning models were able to generate a classification model with a very high accuracy. The most accurate model was the decision tree model, which had an accuracy of over 99.54%, which aligns with our hypothesis that clustering algorithms would work the best since similar values were always clustered together. The other models still provided high accuracy scores, the KNN model produced an accuracy of over 78.90%, when the number of neighbors was set to 4, which indicated that any more neighbors overprioritized exploration over exploitation, and thus reduced the accuracy significantly.

Even though the MLP Classification model was the most time-consuming, it eventually converged with an accuracy of 75.05% at 79 iterations, although the accuracy fluctuated widely whenever the hyperparameter was increased by one iteration.

Conclusion

My research indicates that there appears to be some promise in using machine learning models to classify solar flares, in particular, the decision tree model. This could signify that machine learning models are more accurate than studying solar cycles, and that my investigation was a success. Although, there will need to be further exploration of these types of models and other ones as more factors need to be considered. For example, the trajectory of the CME should be taken into account to investigate the possibility of colliding with the earth's atmosphere.

One model that should be taken into consideration for further investigation would be to use the Qiskit Machine Learning Library for more complex analysis of the underlying factors of the exact photonic interactions. This would also provide some degree of novelty and originality in my research as it provides cutting edge quantum neural networks to analyze the data. Furthermore, the reason the models returned such a high accuracy was because the nature of radial classification was simple enough based on an analysis of the dataset. A more challenging classification approach would be to use an actual scientific classification model that classifies flares by strength.

The classification model used by NASA classifies flares with magnitudes A , B, C, M and X, where each letter represents a tenfold increase in energy output and X class flares being the most dangerous.

Acknowledgments

This project could not have been completed without the mentorship of Aidan Donaghey (MS, Yale University). I also offer my sincere thanks to the INSPIRIT AI Team for giving me this opportunity.

References

Samaha, Kheirallah. “Solar Flares from RHESSI Mission.” *Kaggle*, 7 Apr. 2021, www.kaggle.com/datasets/khsamaha/solar-flares-rhessi.

https://www.space.com/the-carrington-event#:~:text=The%20Carrington%20Event%20was%20a,on%20the%20solar%20disk%20grew.

Dobrijevic, Daisy. “Coronal Mass Ejections: What Are They and How Do They Form?” *Space.com*, Space, 26 May 2022, www.space.com/coronal-mass-ejections-cme.

Fletcher, L., et al. “An Observational Overview of Solar Flares - Space Science Reviews.” *SpringerLink*, Springer Netherlands, 11 Aug. 2011, link.springer.com/article/10.1007/s11214-010-9701-8.

“Learn.” *Scikit*, scikit-learn.org/stable/.