SOLAR FLARES PREDICTION [Support Vector Machines] – AI

**Awais Hussain / UB: 19009475**

# *Abstract*

The purpose of this report is to develop and explore methods of A.I that can be implemented to predict the rate and currents of solar flares and why this is helpful for us. This will be done by development of an AI agent, along with seeking possible improvements for future developing.

# **Contents**

# Introduction

# Background

# Methodology and Data

# Improvements and Evaluation with Criticism

# Conclusion and Future Developing

# Appendages

# **Introduction (1)**

This report will show investigations on how Support Vector Machines are systematically used for classification to discover the occurrences of solar flares from the sun. Accuracy and measurements of program execution will be presented to display how effectively we can approach this situation, showing if the results can be valuable, or if there is to be potential improvements to support the industry.

A Solar flare is magnetic energy that has been built up in the Sun that is unexpectedly released. The radiation that left the sun was in the spectrum of radio waves (example: x-rays and gamma), the researchers that study this field classified those radio waves from solar flares based on their power, to determine if it would potentially harm or damage Planet Earth. The classes of those waves are called X, M,C,B and A-class flares. The X-class flares can reach the earth and cause radiation storms (Qahwaji, 2021). The quantity of the energy that was released is so powerful that it could give energy and power to our planet for 10 million years.

# **Background (2)**

Being able to detect Solar Flares is key to sustaining a healthy lifestyle for humans on Earth, issues with Solar Flares ruining transformers, power grids and other electrical equipment in towns and cities, overloading them with massive electromagnetic energy thus shutting down power, causing potentially catastrophic issues for people, plane companies have to be assured that solar flares won’t happen whilst they’re in flight or else the plane will malfunction and crash, also, these solar flares are radioactive and potentially dangerous by giving radiation to everything that's exposed to the Sun, this makes Solar Flares especially dangerous to pregnant women and can give increase risk of miscarriages. *(Rakshit, 2021).*

With current technology, we are still rather slow in detecting Solar Flares, we can judge how big a storm will be within 60-90 minutes before the Storm will hit Earth *(Rakshit, 2021)*, however this is improving with the U.K and U.S.A putting more effort and resources into space research and developing technologies that can aid in this matter.

The approach that will help aid in these predictions, such as the Support Vector Machines model, which is fast and dependable even with a limited amount of data to analyse *(Stecanella, 2017)*. This will be a supervised machine learning model using classification algorithms for two-group classification problems. Once the Support Vector Machine has been given their sets of training data, they will be able to categorise new text. With this, the model will have a higher speed and better performance even with a limited number of samples. *(Stecanella, 2017).*

# **Methodology and Data (3)**

**Pre-processing and Data Gathering**

The data set from the Machine Learning Repository is used, named ‘Solar Flare Data Set’. This dataset has been used and gathered throughout several science departments, which consists of over 1000 scans to detect presences of solar flares based on the attributes on the list shown below.

|  |  |
| --- | --- |
| 1. modified Zurich class | 1. largest spot size |
| 1. spot distribution | 1. Activity |
| 1. Evolution | 1. Previous 24-hour flare activity code |
| 1. Historically-complex | 1. Area |
| 1. Did region become historically complex on this pass across the sun's disk? | 1. Area of the largest spot |

These are contained within an excel spreadsheet with a large quantity of data. Each scan represents results for all classes (X, M or C), meaning the occurrences of solar flares are viewable, along with what type of solar flare class, since classes will determine their intensities (X being the most dangerous, whilst C being the least threatening).

The flowchart below describes how the program is constructed:

Diagram

Description automatically generated

**Methodology and Development**

When developing the program, sklearn was heavily utilized as it was efficient to use their pre-trained model, Support Vector Machines. This helps processing data of this size, but also works with numbers well when it comes to predictions and classification. Support Vector Machines can predict more than a 2-dimensional space (like with binary which consists of 0 and 1), as SVMs increase their dimensions on outcomes based on how many numbers of outcomes it has. With use of the ‘Kernel Trick’ in SVM, we are able to sidestep expensive calculations of new dimensions, allowing it to calculate results with less frequent issues. (*Stecanella, 2017*).

Before training, null values were evaluated and values of the data had to be converted to integers or floats from string values, as the pre-built SVM can only work with numbers. To combat this issue, replacements specifically within columns were made.

Table

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidence

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

The data set was then split into X\_train, Y\_train, X\_test, Y\_test. Both X values will contain the predictor variables whilst Y will hold the result variables. The industry standard of ratio was chosen here, giving the model an 80:20, [Training : Test] ratio split.

Text, letter

Description automatically generated

The training variables are fit to the SVM AI model (with use of the kernel trick) so that it can learn from this data and use it for any future data given.

Graphical user interface, text, application

Description automatically generated

After training, The prediction score for training data was 83%, while test accuracy was 82%.

*This image shows how the AI model is set to predict training and testing data, giving a prediction score.*

Text

Description automatically generated

# **Evaluation, criticism, and potential improvements (4)**

**Evaluation**

Once displaying the results on a confusion matrix (for C-Class), the AI model predicted many 0 values to be correct and only one 1 value, along with it miscalculating numbers 2-5. The reason for this is because there are many 0 values within the data so it will be accurate at trying to predict, since most consistent values are 0. This had a heavy 0 bias so I decided to discard the 0 values temporarily to see how it would affect prediction.

*Confusion Matrix C-Class [With 0 values Without 0 values*

Calendar

Description automatically generated with medium confidenceCalendar

Description automatically generated

*Confusion Matrix M-Class With 0 values Without 0 values*

A picture containing graphical user interface

Description automatically generatedText, calendar

Description automatically generated

*Confusion Matrix X-Class With 0 values Without 0 values*

A picture containing graphical user interface

Description automatically generatedChart

Description automatically generated with low confidence

From the confusion matrices, it is a lot simpler to view that the data set is biased to a specific type of outcome, which is 0. This is due to majority of the results ending up with a 0 in all 3 classes, along with M and X-class not having results close to C-class, since C-class has the most common occurring flares. The program is mostly predicting 0, which ends up making errors due to how many results in the program are 0. By removing the 0 values, the data is still inaccurate, as there is not enough data provided for the program to make a clear prediction for each class.

A screenshot of a computer

Description automatically generated with low confidence

There are too many 0 values to assist the model to make predictions outside of 0. Therefore, the program will consistently get a good accuracy since majority of the data set consists of 0. The removal of the 0 value was done to see how it would affect the prediction outcome.

A picture containing table

Description automatically generated

Table

Description automatically generated*The table on the right show’s execution run times for the program and how efficient it is:*

*The next page onward will display accuracies of all 3 classes (X, M, C classes) and how the removal of 0 will affect the program and its bias.*

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated

Table, calendar

Description automatically generated

**Improvements and Evaluation with Criticism**

1. Confusion Matrix did not accurately align most results outside of 0. Lacked the importance of true positives and false negatives.

**Improve:** Minimise the bias in the data and give more examples of other predictions.

1. The reason M-class started to have a decreased testing average score, was due to there being not enough values to support a good accuracy, since the program is going to commonly seek out anything besides 0, which again is not enough data from the UCI data set.

**Improve:** Use a data set that is more consistent in its result attributes.

1. Removing 0 for X-class was clearly inefficient, as it had only 5 outcomes not including 0, meaning it would always get a 100% accuracy from having very little data.

A picture containing table

Description automatically generated

**Improve:** Add extra data for X-class by (another data set or) own research

1. Removing 0 was not an entirely accurate solution of showing how the program should make new predictions and tests, as it was just shown to compare situations of when 0 was present or not.

**Improve:** Adjust the weighting percentage to get a fairer result.

# **Conclusion and Future Developing (5)**

After the testing of this program, it is highly advised that a data set with a good amount of outcome varieties will determine a better prediction and accuracy, as well as having a more expansive amount of outcomes. Weighting the data set to give it less frequent answers of what is common will help reduce a bias within training.

The outcome of these tests, however, have proven to us that out of the three classes, C-Class flares occur the most, whilst X-Class are the least occurring and M-Class being in the middle. With this, we are more aware of how often flares are happening and the severity of them to our civilisation, thus it is advised to understand the importance of preventing these issues and to develop better technologies to protect ourselves from them.

With more time, we would have used a variety of data sets to show a non-biased result as well as create an A.I with different approaches in an attempt to have a higher accuracy in detection of these Solar Flares.

## Bibliography

2021. *Strong solar flares and how they impact Earth. Mint.* Available at: <https://www.livemint.com/science/news/strong-solar-flares-emitted-from-sun-towards-earth-how-it-may-impact-us-11635574893358.html> (Accessed: 17/12/2021)

2013. Fox F, *Impacts of Strong Solar Flares, NASA.* Available at:<https://www.nasa.gov/mission_pages/sunearth/news/flare-impacts.html> (Accessed: 17/12/2021)

2015. McDonald J. *The Threat of Solar Flares. The Daily Jstor.* Available at: <https://daily.jstor.org/dont-underestimate-threat-solar-flares/> (Accessed: 17/12/2021)

2021. Rakshit D. *Solar Storms are Back. How do they Affect Life on Earth. theswaddle.* Available at: <https://theswaddle.com/solar-storms-are-back-how-do-they-affect-life-on-earth/> (Accessed: 11/01/2022)

2017. Stecanella B. *Introduction to Support Vector Machines. monkeylearn.* Available at: <https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/> (Accessed: 11/01/2022)

2017. Stecanella B. “*Support Vector Machines (SVM) Algorithm Explained”. monkeylearn.* Available at: <https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>

Accessed: 22/06/2017)

2021. Kabed A, Qahwaji R. Abed A. *The automated prediction of solar flares from SDO images using deep learning. Science Direct.* Available at: <https://www.sciencedirect.com/science/article/pii/S0273117721000971?via%3Dihub#b0165>

(Accessed: 11/01/2022)

# **Appendages (6)**

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated