**Introduction**

Solar activity can have deleterious effects on both terrestrial and space-based systems. Powerful eruptions on the surface of the sun known as solar flares eject high energy charged particles. Solar flares are categorized as A, B, C, M, or X on a logarithmic scale according to their peak x-ray flux (W/m2) in the 0.1 - 0.8 nm wavelength range. M and X class flares are the most powerful and pose the largest threat to infrastructure. Charged particles from such events can interact with Earth’s magnetic field and cause slowly varying currents in the upper atmosphere (ionosphere). Currents in the ionosphere can, in turn, induce damaging direct currents in man-made systems such as transcontinental pipelines and electricity transmission networks. The ability to predict solar flares from observable solar features, such as sunspots, would allow protective measures to be taken on critical systems prior to the arrival of the flare.

It has been shown that sunspot activity is closely related to solar flare emission [1]. Sunspots are regions on the sun’s photosphere where intense magnetic fields cool the surrounding plasma to temperatures that are several thousand Kelvin less than the surrounding photosphere gasses. As a result of this temperature gradient sunspots appear as dark spots. Sunspots consist of a dark core region, the *umbra*, and a surrounding less dark area of higher temperature, the *penumbra*. The international solar physics community classifies sunspot types using the three component McIntosh classification scheme [2]. McIntosh classifications consist of a three letter code typically abbreviated as *Zpc* where *Z* describes the Zurich group type, *p* represents the penumbra type, and *c* represents spot compactness. Examples of the McIntosh classification scheme are shown in figure 1 and allowed classifications are tabulated in table 1. Sunspot locations are given by their National Oceanographic and Atmospheric Administration (NOAA) active region number. Sunspots must be observed by two separate observatories to be assigned a NOAA active region number. This work attempted to predict solar flare events given a sunspot group type.

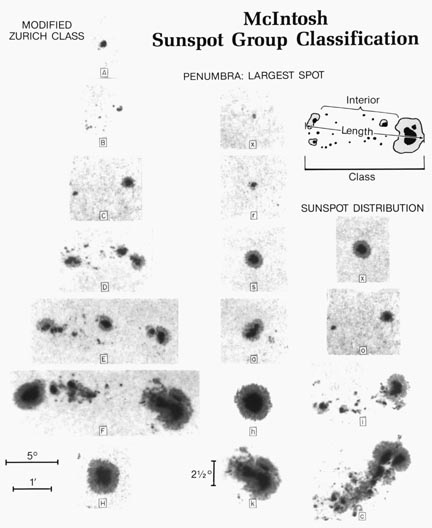


Figure 1: The 3-component McIntosh classification, with examples of each category [2].

|  |  |  |  |
| --- | --- | --- | --- |
| Group Type | Penumbra of Largest Spot | Spot Compactness | Number of Types |
| A | x | x | 1 |
| B | x | o, i | 2 |
| C | r, s, a, h, k | o, i | 10 |
| D, E, F | r | o, i | 6 |
| D, E, F | s, a, h, k | o, i, c | 36 |
| H | r, s, a, h, k | x | 5 |

Table 1: Allowed types of groups in McIntosh sunspot classification. There are a total of 60 allowed classifications.

**Data Preparation**

Data for initial model development was provided by the University of California Irvine (UCI) machine learning repository [3]. The data set contained 1066 observations of NOAA active regions measured from August 8, 1978 to December 23, 1978. Thirteen predictors were present (see data dictionary). The UCI data is a well ordered data set used for teaching, and very little preprocessing was needed. For each instance the number of C, M, and X type solar flares were summed and added to a new response column. The values of this column were then mapped to a binary response, 0 if no solar flare was observed and 1 otherwise.

A second set of data for model evaluation was provided by the NOAA National Geophysical Data Center (NGDC). The Solar-Terrestrial Physics (STP) division at NGDC has made Mount Wilson Observatory sunspot group data and Geostationary Operational Environmental Satellite (GOES) solar flare event data available on their ftp site. A Python script (goes\_sunspot\_parser.py) was used to gather a larger set of data from 1982 to 2014. X-ray data consisted of measurement date, x-ray flare class, and NOAA active region number where as sunspot group data consisted of measurement date, the three component McIntosh classification, and NOAA active region number. Two separate data frames containing sunspot group data and x-ray event data were created.

Raw x-ray event and sunspot group data were cleaned using a Python script entitled sun\_x\_ray.py. It was decided to only investigate M and X class solar flares, as these events pose the most severe threat to infrastructure. The sunspot data was parsed so that only sunspot observations with NOAA active region numbers present within the x-ray event data frame were kept. A sunspot group tended to have multiple observations over the course of a day. Each solar flare observation, however, was unique. A data frame of solar flare measurements and associated sunspot observations was created by matching sunspot groups and flare events using the NOAA active region number. If a flare and a sunspot group shared a common NOAA active region number, the time difference between the flare event and each sunspot measurement in the group was calculated. Sunspots that are observed within six hours of a flare event were considered to be associated with the flare. If multiple sunspot observations were made within six hours of a flare measurement, only the observation with the minimum time difference was considered to be associated.

The associated sunspot group observations were merged with the appropriate solar flare observations and written to a data frame, sunspot groups that were not associated with a solar flare observation were written to a separate data frame. The final products were two data frames, the first consisting of the associated solar flare class and the three component McIntosh sunspot classification (Table 2), and the second consisting of the three McIntosh labels for unassociated sunspots.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Zurich | Penumbra | Compactness | Flare Class |
| 1981-12-28 20:28:00 | D | r | o | M |
| 1982-01-08 13:54:00 | E | k | o | M |

Table 2: Associated solar flare/sunspot data example.

**Analysis**

The goal of this work was to predict the occurrence of a solar flare given information related to sunspot observations. The UCI data set was used to test three types of machine learning methods and subsequently build a predictive model. One binary response vector was used throughout; if any type of solar flare was observed for an instance the response was 1, 0 was used if no flares were observed. Due to the binary nature of the response logistic regression, decision trees, and naïve Bayes methods were investigated. The use of McIntosh sunspot classification values was kept constant throughout all three methods. Recall (sensitivity) was considered to be the most important scoring metric for the analysis

Logistic regression and decision trees were used to classify each instance into flaring or non-flaring. To implement both methods, columns of dummy variables were created for the McIntosh (Zpc) classification values, thereby adding thirteen new predictors to the data set. The models were tuned by the addition of extra predictors. A ten fold cross validation method was used to score the models; both the receiver operator characteristic area under the curve (ROC AUC) and the sensitivity were used as scoring metrics. The best performing logistic regression model utilized seventeen predictors: thirteen McIntosh dummy variables, activity, previous 24 hour activity, and historic complexity. This model yielded an ROC AUC score of 0.796, however, sensitivity was only 0.279. A decision tree performed worse than logistic regression. The most promising decision tree featured a maximum depth of four and relied solely on the thirteen McIntosh dummy variables. The decision tree ROC AUC and sensitivity metrics were 0.741 and 0.221 respectively. Both models exhibited ROC AUC values that appeared promising, however, this was found to be related to the large number on non-flaring events (81% of instances did not flare). The sensitivity values were low indicating that each model correctly predicts a flaring event less than 30% of the time.

The best performing predictive model utilized the naïve Bayes classifier. For this model a new predictor was created by concatenating the Zurich, penumbra, and compactness columns into a single Zpc classification column. This new feature was entirely text based and was used to create a predictor document term matrix. A ten fold cross validation was used to score this model. The ROC AUC and sensitivity values were 0.751 and 0.513 respectively. This sensitivity value is still low, but the naïve Bayes model performs well enough that it could be included in an ensemble method.

**Conclusion**

The application of machine learning to solar flare prediction is promising, however, the use McIntosh sun spot classifications as a predictor is inherently flawed. The best predictive model generated for this work only predicted flare events correctly 51.3% of the time. The classification of sun spots is not standardized and done entirely by humans who will be directed their unique biases when assigning Zpc labels. Future models should include quantitative predictors, such as magnetic field strength, in addition McIntosh labels.

**References**

1. H. Zinn and M.A. Liggett, *Sol. Phys.* **113**, 267 (1987)
2. P.S. McIntosh, *Sol. Phys.* **125**, 251 (1990)

1. <https://archive.ics.uci.edu/ml/datasets/Solar+Flare>