Space Weather Analytics Dataset for Solar Flare Predictions (SWAN-SF)

Berkay Aydin
Oct 2, 2019
DMLab Workshop
Atlanta, GA
Buzzwords...

- Artificial Intelligence
- Machine Learning
- Data Mining
- Knowledge Discovery
- Data Science
- Deep Learning
- Data Analytics

Source: https://www.houseofbots.com/
Hunger

All the learning models are data hungry

Me WANTS THE DATA
**The FOUR V's of Big Data**

- **Volume**: It's estimated that 2.5 quintillion bytes (2.3 trillion gigabytes) of data are created each day. By 2020, 40 zettabytes (45 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005.
- **Variety**: As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (161 billion gigabytes). By 2014, it's anticipated there will be 420 million wearable, wireless health monitors. 4 billion hours of video are watched on YouTube each month. 400 million tweets are sent per day by about 200 million monthly active users.
- **Velocity**: Most companies in the U.S. have at least 100 terabytes (100,000 gigabytes) of data stored. The New York Stock Exchange captures 1 TB of trade information during each trading session. Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure. By 2016, it's projected there will be 18.9 billion network connections — almost 2.5 connections per person on earth.
- **Veracity**: According to Hadoop, 70% of large data sets are inaccurate. In one survey, 73% of IT leaders said their data quality is incomplete. 1 in 3 business leaders don’t trust the information they use to make decisions. 75% of survey respondents reported that 27% of the data they use is inaccurate or uncertain.

**Volume of Data**

- World population: 7 billion
- 6 billion people have cell phones

**Variety of Data**

- 30 billion pieces of content are shared on Facebook every month
- 400 million tweets are sent per day by about 200 million monthly active users

**Velocity of Data**

- The New York Stock Exchange captures 1 TB of trade information during each trading session
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure

**Veracity of Data**

- 70% of large data sets are inaccurate
- 73% of IT leaders said their data quality is incomplete
- 27% of the data they use is inaccurate or uncertain

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, WEFTEC, GGS
Two common misconceptions

● Algorithmic development is the only driving factor behind the data driven discoveries
  ○ The groundwork for transformational advancements in data science is laid by the algorithmic development; however, major breakthroughs are only realized by the availability of high quality data.

● More data is better
  ○ Quality of the data is more impactful than the quantity
  ○ Adding spurious data to training sets will generally result in decreased accuracy performance
Data Quality

● The driver for data quality problems
  ○ Lack of data strategy: very little thought is given to the expense of collecting/storing data

● Components of data quality
  ○ Completeness: No gaps in data
  ○ Consistency: versions and data types are compatible
  ○ Accuracy: is the data correct?
  ○ Validity: Can we validate the correctness of data?
  ○ Timeliness: Are we getting the critical data on time?

● Why does it matter? Models are affected
**Benchmark**

*from Merriam-webster*

1. Something that serves as a standard by which others may be measured or judged

2. A point of reference from which measurements may be made

3. A standardized problem or test that serves as a basis for evaluation or comparison (as of computer system performance)
Benchmark Dataset

...datasets used for benchmarking a particular class of algorithms and systems

- Standardized testing/validation
- Common ground for comparison
- Ease of access
- Usability
Data & Prediction

How can we decide which model is better than the other?

accuracy, f-score, AUC, skill scores, ...?
Data & Prediction

2 different datasets

2 forecast models

model A

model B
Data & Prediction

- many small decisions...

- Representative of an entire Solar cycle?
- Multi-class or binary dataset?
- Cut-off for positive and negative classes: C5.9? M1.0? X1.0…?
- Equally difficult training and testing sets?
- Under-sampling? Over-sampling?
- Balanced or imbalanced?
- Which normalization methods?

- 2 different datasets
- 2 forecast models

- model A
- model B
Usefulness

- Creating datasets for learning models involves
  - Data wrangling
  - Cleaning
  - Pre-processing
  - Normalization
  - … “data janitor work”
- Least enjoyable part

Source: CrowdFlower Data Science Report 2016
Usefulness

● What if datasets created in advance?
● DS/ML/DM/DL/KDD community kept doing this
  ○ OpenNeuro (MRI, MEG, EEG) – https://openneuro.org/
  ○ Kaggle Datasets – https://www.kaggle.com/datasets
Standardization

● Using the same dataset (this is a challenge!)
  ○ Preferably under the same validation settings (this is not a challenge)
    ■ Same accuracy measurements, same/similar sampling strategies
  ○ We cannot force anyone to use any dataset
    ■ But, we can *kindly* ask them to predict the same set of events
    ■ Under (at least) similar circumstances
  ○ Great SWx Bake Off

● Fair comparisons

● Preferred accuracy measurements
  ○ Probability of detection
  ○ False alarm rate
  ○ Skill scores
Standardization - Challenges

● Living datasets
  ○ Solution: Updates (e.g., annual)

● Accessibility
  ○ Shared with standard formats
  ○ Common access paths
  ○ Reproducible data science

● Code
  ○ To manipulate data
  ○ Open source platforms
  ○ Reproducibility
Outreach with Benchmark Datasets

- Community development
  - Many people working on cat pictures
- Accessible high-quality data = less data janitor work
- Showcase the problem
  - Attract interest from interested data scientists
  - Interdisciplinary and cross-disciplinary research
SWAN Benchmark Datasets

- **Space Weather Analytics (SWAN) Benchmark Datasets**
  - Solar Flares (current)
  - CMEs & SEPs will come next

- **Rationale: Data sources used for SWx forecasts are similar**
  - Magnetic field data
  - Sunspots
  - (Rarely) AIA

- **Solution: standardization of the validation**

- **Currently available through Kaggle**
  - via a Big Data Cup Challenge
Challenges for SWx Forecasting Datasets

● Rare events
  ○ Class imbalance – extreme

● Data integration
  ○ Heterogeneous data sources
    ■ X-ray time series
    ■ Source active region properties – from HMI vs NOAA ARs
  ○ Formats and code (open source platforms vs IDL)
  ○ Spatial and temporal resolution

● Data coverage
  ○ Not all data is available for all the events
Common data issues for SWx forecasting

- **Senseless random undersampling**
  - Rare events -> class imbalance -> undersampling
  - Keep all the positive data instances
    - E.g., major flares or erupting flares
  - Randomly undersample without considering underlying data distribution
    - E.g., 1:1 balance for flare prediction
    - Representative samples

- 10 X-class
- 90 M-class
- 15 C-class
- 10 B-class
- 75 Flare-Quiet
Common data issues for SWx forecasting

- No time segmentation of datasets
  - Randomly assigning data instances to training and testing subsets without considering the temporal dependencies

- The most similar data instance for a given instance
  - Usually the subsequent or consequent
  - If temporal neighbors are placed in training and testing datasets
    - Models will not learn
    - They will simply memorize (cannot generalize).
≈ using training data for testing
Solution

- Separate dataset into temporally non-overlapping time intervals and use the data instances from these as a block
Common data issues for SWx forecasting

- No cross validation
  - K-fold cross-validation
  - Requires more experiments
  - Necessary for verifying models ability under different parts of solar cycle or different solar cycles
Integrating Data for SWAN

1. Verifying GOES Flares
2. Magnetic Field Parameters
3. Integrating Flares
4. Dataset Generation
Verifying GOES Flares

- **GOES Flare List** (Primary source)
  - Temporal Attributes: Start - Peak - End Times
  - Spatial Attributes: Coordinates (Explicit) + NOAA AR (Implicit)
  - Magnitude (e.g. M2.3)

- **Secondary sources**
  - Flares from SSW Latest Events (AIA)
  - Hinode-XRT Flare Catalog

- **NOAA ARs**
  - Trajectories of daily AR event reports
Verifying GOES Flares

Step 1: GOES Flares usually do not have explicit locations (many have NOAA ARs). Approximate the missing explicit locations of flares using NOAA AR reports.

Example flares with no location from AR# 12567:

M5.0 Flare: SOL2016-07-23T01:46:00L100C090
Approximate loc: (67.3, 6)

M7.6 Flare: SOL2016-07-23T05:00:00L098C090
Approximate loc: (69.1, 6)
Verifying GOES Flares

Step 2: GOES Flares Cross-checking

Step 2.1: For each GOES Flare –
Find a co-occurring SSW flare (time+magnitude)
Find a co-occurring XRT flare (time+magnitude)

M5.0 Flare from GOES:
AR# 12567
SOL2016-07-23T01:46:00L100C090
Approximate loc: (67.3, 6)
Corresponding XRT and SSW Flare Reports

147410  2016/07/23 01:46 - 02:11 - 02:23UT  N05W73  M5.0

Flare Locator Image
Verifying GOES Flares

Step 2: GOES Flares Cross-checking

Step 2.2: For every triplet of GOES-SSW-XRT flares

If GOES location is within 275 arcsec of SSW or XRT location → **Primary-Verified**

Else if SSW location is within 275 arcsec XRT location → **Secondary-Verified**

Else → **Non-Verified**
Figure A.2: Reported coordinates of the X3.3 flare occurred on 2013-11-05 at 22:12:00 from GOES, SSW Latest Events and Hinode-XRT. The Euclidean distances between these coordinates are recorded for distance-based verification. Coordinates are reported in Helioprojective Coordinate System. $d_{GS}$ is the distance between GOES and SSW, $d_{GX}$ is the distance between GOES and XRT, and $d_{SX}$ is the distance between SSW and XRT.
Verifying GOES Flares

Step 2: GOES Flares Cross-checking

Step 2.3: For each Secondary-Verified flare
Augment SSW Location

Example:
M1.7 Flare at 2012-11-20T12:36:00
Location: Undetermined
AR# Unknown

SSW Location augmented (-20, 6)
Resulting List of Flares

![Bar chart showing the count of flares by class and verification status.]

- **Primary-verified Flares**
  - A or B: 3,929
  - C: 6,225
  - M: 56
  - X: 50

- **Secondary-verified Flares**
  - A or B: 1,376
  - C: 1,331
  - M: 12

- **Non-verified Flares**
  - A or B: 550
  - C: 198
  - M: 674

Legend:
- Blue: Primary-verified Flares
- Orange: Secondary-verified Flares
- Green: Non-verified Flares
Overview

- NOAA/GOES Flare Reports
- SSW Latest Events Flare Reports
- Hinode/XRT Flare Reports
- NOAA/SRS Daily Active Region Reports
- SDO/HMI SHARPs
- NOAA/GOES X-ray Flux

Active Region Centroid Augmentation

Flare Matching on Time and Magnitude
Distance-based Verification
Secondary Location Augmentation

Cleaned and cross-checked flare reports

Magnetic Field Parameter Generation
NOAA AR Number Integration
X-ray Flux Integration
Flare History Parameter Integration

Machine-learning-ready Dataset Creation
Partitioning, Slicing & Labeling, Undersampling

MVTS Dataset for Space Weather Analytics
Magnetic Field Parameters

- Based on SHARPs and SHARP keywords†
  - Same identifier [HARPNUM]
- Multivariate time series
  - Each parameter is a quantity of magnetic field
    - E.g. Unsigned flux, Mean current helicity, magnitude of Lorentz force etc.
  - Sampled with 12-minute cadence

Magnetic Field Parameters – Pipeline

- A parallel computation pipeline
- From Definitive SHARPs -- Cylindrical Equal Area (CEA)
  - Data series: hmi.Sharp_cea_720s
- Computes 24 magnetic field parameter from vector magnetograms
  - $B_r$
  - $B_\phi$
  - $B_\theta$
  - + BITMAP (AR boundary) and CONF_DISAMMB (confidence map of MF disambiguity)
  - Error files omitted for computational and storage efficiency
NOAA AR Number & X-Ray Flux Integration

- We match the HARP segments’ locations with NOAA AR locations
  - Already provided in the keywords, but probably for NRT series
  - However, we have found some anomalies and fixed them (both NOAA AR and SHARPs)

- Also included global X-ray flux
  - Cleaned using primary and secondary GOES sources
  - Quality flag is available
Integrating Flares

Cleaned flares integrated based on associated NOAA AR#

Binary TS for B, C, M and X-class flares
Overview
ML-ready Datasets

1. Partitioning
2. Slicing
3. Labeling
Partitioning

- Rare event prediction
- Primary focus is on the task
  - Predicting major flares (>M1.0)
- Time segmented stratification
- Equally distribute the number of major flares to partitions
Slicing & Labeling

- Sliding window approach

Using a sliding window, we get a slice of MVTS.

The slices are labeled using the flare with max
magnitude within the next 24 hours.

Label: X1.2
Observation Window: $T_{obs} = 8\tau$
Latency: $L = 4\tau$
Prediction Window: $T_{pred} = 6\tau$

Label is set as the magnitude of largest flare during the prediction window (e.g. M1.6, C7.0) or FQ if no flares are reported.
Extending the dataset

*Horizontal extensions*

- Adding new parameters to the dataset

**Examples**

- Magnetic Field Parameters from NRT series
- Number of close-by sigmoids to SHARP bounding box

MVTS Parameters Derived from a SHARP collection
Extending the dataset

Vertical extensions

- Adding new target SW events (to predict)

Examples

- CMEs, SEPs (>10 MeV, >100 MeV)
- Filament eruptions
Thanks

Questions?