

CS 780 MMs Class Project

Preliminary Report

Vega Group: Daroc Alden, Samantha Piatt, and Jeremy Walker

April 12, 2018

Motivation

This project is intended to explore options for automating a key part of the ongoing Magnetospheric Multiscale Mission - the selection of which detailed datapoints ought to be downloaded from the satellite. This job is currently done by a Scientist in the Loop, who must spend time each day evaluating the data observed by the satellite to decide how to spend bandwidth resources.

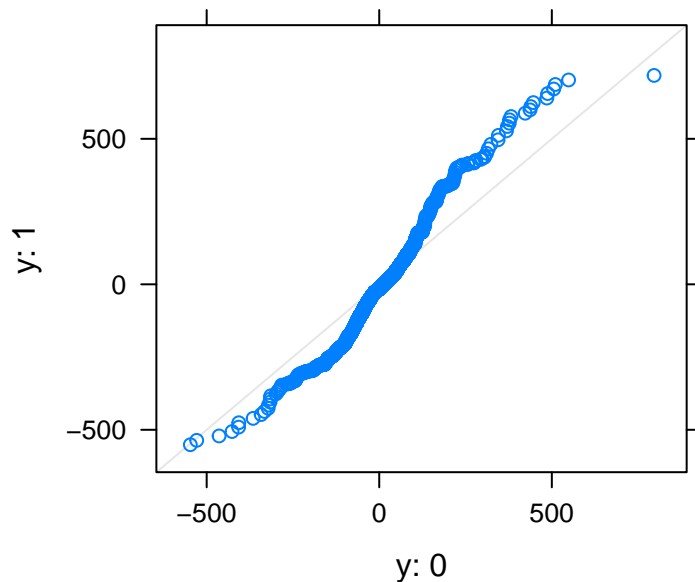
Automating the selection of interesting data would free up valuable human time on the project. To do this, we have explored several methods for determining what data points might be interesting, as detailed below.

Related work

(papers describing machine learning methods or their applications)

Evaluation criteria

We found that the data was not identically and independently distributed, as demonstrated by the exemplary QQ plot below.



In fact the data is a time series. Therefore, we could not use the usual method of keeping some test data out to evaluate the performance of our model, because it would have left discontinuities in the data used for training our model.

Therefore, we are using one of the two data sets as a validation set, while using the other data set in full to train our model. We hypothesise that adding time-delayed factors (i.e. factors which contain values from previous samples) will significantly improve the performance of many machine learning models on the validation set. See below for details.

Methods

We compared our time series models against the basic feature set with no consideration for time, using multiple machine learning methods.

Our first time series model, which we will call time shifting, involves adding n columns to the data that contain each feature shifted $n = 5$ points in the past. The second time series model, which we will call time difference, involves adding a column which takes the average difference between the current data point x and the past $n = 20$ data points.

Logistic Regression

```
## Basic MMS data
## -----
## Test set of 302 SITL points - Found: 143 Missed: 159
## Training set of 932 SITL points - Found: 444 Missed: 488
## Total set of 1234 SITL points - Found: 589 Missed: 645
## Classification Error - Test: 0.1367742 Training: 0.1399885 Total: 0.1387544

## Time Shift data
## -----
## Test set of 296 SITL points - Found: 136 Missed: 160
## Training set of 938 SITL points - Found: 494 Missed: 444
## Total set of 1234 SITL points - Found: 629 Missed: 605
## Classification Error - Test: 0.1377529 Training: 0.1274214 Total: 0.1302195

## Time Difference data
## -----
## Test set of 302 SITL points - Found: 154 Missed: 148
## Training set of 932 SITL points - Found: 475 Missed: 457
## Total set of 1234 SITL points - Found: 631 Missed: 603
## Classification Error - Test: 0.1273118 Training: 0.1310958 Total: 0.1297193
```

Bagging

Using 16 splits (number of features).

```
## Basic MMS data Set
## -----
## Test set of 302 SITL points - Found: 236 Missed: 66
## Training set of 932 SITL points - Found: 932 Missed: 0
## Total set of 1234 SITL points - Found: 1168 Missed: 66
## Classification Error - Test: 0.05677419 Training: 0 Total: 0.01419813
```

```

## Time Shift data
## -----
## Test set of 296 SITL points - Found: 229 Missed: 67
## Training set of 938 SITL points - Found: 938 Missed: 0
## Total set of 1234 SITL points - Found: 1167 Missed: 67
## Classification Error - Test: 0.05768403 Training: 0 Total: 0.01442101

## Time Difference data
## -----
## Test set of 302 SITL points - Found: 241 Missed: 61
## Training set of 932 SITL points - Found: 932 Missed: 0
## Total set of 1234 SITL points - Found: 1173 Missed: 61
## Classification Error - Test: 0.05247312 Training: 0 Total: 0.01312251

```

Random Forests

Using 4 splits (square root of number of features).

```

## Basic MMS data Set
## -----
## Test set of 302 SITL points - Found: 239 Missed: 63
## Training set of 932 SITL points - Found: 932 Missed: 0
## Total set of 1234 SITL points - Found: 1170 Missed: 64
## Classification Error - Test: 0.05419355 Training: 0 Total: 0.01376788

## Time Shift data
## -----
## Test set of 296 SITL points - Found: 229 Missed: 67
## Training set of 938 SITL points - Found: 938 Missed: 0
## Total set of 1234 SITL points - Found: 1167 Missed: 67
## Classification Error - Test: 0.05768403 Training: 0 Total: 0.01442101

## Time Difference data
## -----
## Test set of 302 SITL points - Found: 239 Missed: 63
## Training set of 932 SITL points - Found: 932 Missed: 0
## Total set of 1234 SITL points - Found: 1171 Missed: 63
## Classification Error - Test: 0.05419355 Training: 0 Total: 0.01355276

```

Radial SVM

```

## Basic MMS data
## -----
## Test set of 932 SITL points - Found: 932 Missed: 0
## Training set of 302 SITL points - Found: 302 Missed: 0
## Total set of 1234 SITL points - Found: 1234 Missed: 0
## Classification Error - Test: 0 Training: 0 Total: 0

## Time Shift data
## -----
## Test set of 938 SITL points - Found: 653 Missed: 285
## Training set of 296 SITL points - Found: 195 Missed: 101
## Total set of 1234 SITL points - Found: 844 Missed: 390
## Classification Error - Test: 0.08179079 Training: 0.08695652 Total: 0.08394318

```

```
## Time Difference data
## -----
## Test set of 932 SITL points - Found: 692 Missed: 240
## Training set of 302 SITL points - Found: 210 Missed: 92
## Total set of 1234 SITL points - Found: 903 Missed: 331
## Classification Error - Test: 0.06884682 Training: 0.07913978 Total: 0.07120577
```

Validation

We're using the second dataset as a validation set. Neither model performs exceptionally well on the validation set. Below is the performance of the Time Difference model on the second dataset.

```
## Radial SVM Time Difference data
## -----
## Total set of 1210 SITL points - Found: 292 Missed: 918
## Classification Error - Total: 0.1625786
## Bagging Time Difference data
## -----
## Total set of 1210 SITL points - Found: 352 Missed: 858
## Classification Error - Total: 0.1519525
```

Recommendations

best method with an estimate of prediction quality

Analysis