

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.

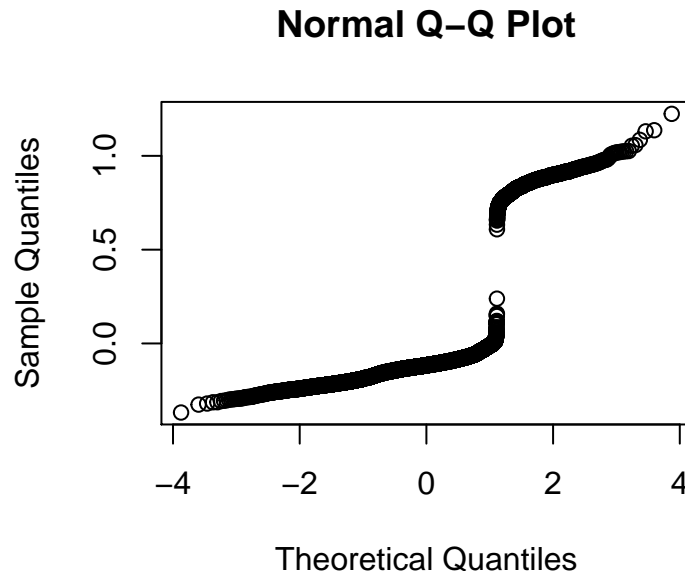
Automating the selection of interesting data would free up valuable human time on the project. To do this, we have explored several methods for determining what data points might be interesting, as detailed below.

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: 157 Missed: 145
## Training set of 932 SITL points - Found: 429 Missed: 503
## Total set of 1234 SITL points - Found: 585 Missed: 649
## Classification Error - Test: 0.1247312 Training: 0.1442915 Total: 0.1396149

## Time Shift data
## -----
## Test set of 313 SITL points - Found: 160 Missed: 153
## Training set of 921 SITL points - Found: 474 Missed: 447
## Total set of 1234 SITL points - Found: 634 Missed: 600
## Classification Error - Test: 0.1317262 Training: 0.1282824 Total: 0.1291433

## Time Difference data
## -----
## Test set of 302 SITL points - Found: 163 Missed: 139
## Training set of 932 SITL points - Found: 474 Missed: 458
## Total set of 1234 SITL points - Found: 638 Missed: 596
## Classification Error - Test: 0.1195699 Training: 0.1313827 Total: 0.1282134
```

Bagging

Using 16 splits (number of features).

```
## Basic MMS data Set
## -----
## Test set of 302 SITL points - Found: 243 Missed: 59
## Training set of 932 SITL points - Found: 932 Missed: 0
## Total set of 1234 SITL points - Found: 1175 Missed: 59
## Classification Error - Test: 0.05075269 Training: 0 Total: 0.01269227

## Time Shift data
## -----
## Test set of 313 SITL points - Found: 248 Missed: 65
## Training set of 921 SITL points - Found: 921 Missed: 0
```

```

## Total set of 1234 SITL points - Found: 1169 Missed: 65
## Classification Error - Test: 0.05596212 Training: 0 Total: 0.01399053

## Time Difference data
## -----
## Test set of 302 SITL points - Found: 251 Missed: 51
## Training set of 932 SITL points - Found: 932 Missed: 0
## Total set of 1234 SITL points - Found: 1183 Missed: 51
## Classification Error - Test: 0.04387097 Training: 0 Total: 0.01097128

```

Random Forests

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

```

## Basic MMS data Set
## -----
## Test set of 302 SITL points - Found: 245 Missed: 57
## Training set of 932 SITL points - Found: 932 Missed: 0
## Total set of 1234 SITL points - Found: 1177 Missed: 57
## Classification Error - Test: 0.04903226 Training: 0 Total: 0.01226202

## Time Shift data
## -----
## Test set of 313 SITL points - Found: 251 Missed: 62
## Training set of 921 SITL points - Found: 921 Missed: 0
## Total set of 1234 SITL points - Found: 1172 Missed: 62
## Classification Error - Test: 0.05337925 Training: 0 Total: 0.01334481

## Time Difference data
## -----
## Test set of 302 SITL points - Found: 254 Missed: 48
## Training set of 932 SITL points - Found: 932 Missed: 0
## Total set of 1234 SITL points - Found: 1186 Missed: 48
## Classification Error - Test: 0.04129032 Training: 0 Total: 0.01032591

```

Radial SVM

```

## Basic MMS data
## -----
## Test set of 932 SITL points - Found: 584 Missed: 348
## Training set of 302 SITL points - Found: 192 Missed: 110
## Total set of 1234 SITL points - Found: 781 Missed: 453
## Classification Error - Test: 0.09982788 Training: 0.09462366 Total: 0.09745079

## Time Shift data
## -----
## Test set of 921 SITL points - Found: 658 Missed: 263
## Training set of 313 SITL points - Found: 202 Missed: 111
## Total set of 1234 SITL points - Found: 860 Missed: 374
## Classification Error - Test: 0.07547711 Training: 0.09556608 Total: 0.08049935

## Time Difference data
## -----
## Test set of 932 SITL points - Found: 690 Missed: 242
## Training set of 302 SITL points - Found: 214 Missed: 88

```

```
## Total set of 1234 SITL points - Found: 908 Missed: 326
## Classification Error - Test: 0.06942054 Training: 0.07569892 Total: 0.07013015
```

Validation

Since, during training, the version of each model using the time difference did significantly better for all models, we will compare those methods' performance on our validation set. Since none of the models perform exceptionally well here, it is likely that they have overfit. We suspect this may be an artifact of the different SITLs on different weeks, or of our comparatively sparse training data. Regardless, the algorithm's performance on this validation set is a good indicator of its likely performance on the hidden test that you have for us, professor.

Also: Since Logistic Regression did so poorly, we've chosen to omit it here for brevity.

Bagging

```
## Bagging Time Difference data
## -----

## Total set of 1210 SITL points - Found: 421 Missed: 789
## Classification Error - Total: 0.1397326
```

Random Forest

```
## Random Forests Time Difference data
## -----

## Total set of 1210 SITL points - Found: 492 Missed: 718
## Classification Error - Total: 0.1271584
```

Radial SVM

```
## Radial SVM Time Difference data
## -----

## Total set of 1210 SITL points - Found: 109 Missed: 1101
## Classification Error - Total: 0.194988
```

Recommendations

Random Forests with Time Difference data did best on our validation set out of all tested methods. With a classification error of 12.73%, they are not as accurate as we might hope, but since we tested each of the methods on a validation set which is similar to the set we will be tested on, we hope that it will have a similar classification error there.

Analysis

Since Random Forests did less well on our training data, but better on our validation set, it seems likely that other methods overfit. This can be solved in a few ways, not least of which will be to acquire more data. Overfitting is evidence that our methods are too flexible - so another way to reduce their flexibility would be

to use bagging or boosting in combination with another method, or to use cross-validation to choose such parameters as the amount of time-shift to use.