Project Report

Group 12

Breif motivation for the project

Evaluation of the quality of the data points gathered by Magnetospheric Multiscale Mission. The spectrum of the data is quite limited due to the orbit of the satellites as well as its buffer storage.

Currently, the current resolution applied and used by scientist is SITL, abbreviation for a scientist in the loop. And the purpose of our project is seeking out a solution that can replace SITL in the future, with a machine learning method which provides a more effective and valuable result.

Related work (papers describing machine learning methods or their applications)

LDA:Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order avoid overfitting ("curse of dimensionality") and also reduce computational costs.It works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis.

Logistic Regression:Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest

Evaluation criteria

1. The valuable data points are those that have been downloaded completely.

Which methods have been tried

Prior to applying machine learning methods, we processed provided data by merging SITL and MMS files with the functions given by professor Petrik. The purpose is putting corresponds, dependent and independent variables into one file that can be used as a unique dataset in our later procedures. The dataset merging process apply to both traning data and test data.

The machine learning methods we applied in our project are LDA, Random Forest, SVM with linear classifier, Logistic Regression, Boosting and Bagging.

LDA

In the LDA model, total SITL points were 1210. The total points matched both scientist pick and machine learning method selected were 163. 1047 points were not selected.The classification error is 0.1854246.

Logistic Regression

In logistic regression, unlick LDA method, the points matched both scientist pick and machine learning methods were 509 which is better than LDA. Meanwhile, the classification error is 0.124 which is better than LDA method.

Random Forest,SVM and bagging

In Random Forest, the points matched both scientist pick and machine learning method selected were 290,the classification error is 0.1629328. The interesting thing is that results of Random Forest, SVM and bagging are same.

Recommended best method with an estimate of prediction quality

The best method with an estimate of prediction quality is Logistic Regression. The result based on the assessment and comparsion of classification error rate, of which we come out with our modules, is 0.124.

The second best method that we got is random forest. The classification error rate is 0.1629328.This result is difference than the previous submission after applied test dataset.

Analysis of the results

In this project, dataset1 is to be tranning set, dataset2 is to be testing set Although the logistic regression method's error is not small, it has the higher matching ratio between scientist selected data and machine learning model selected data.

Based on this model, there are total 493 points used to find the proper time period to download data. They are list as follow: The time gap is setted to be 2 mins. Every two sets of time slot are one group which is estimated to be the good time to download data.

```
## [1] "2015-12-04 00:50:19.259 UTC"
## [2] "2015-12-04 00:52:20.760 UTC"
## [3] "2015-12-04 01:48:58.286 UTC"
## [4] "2015-12-04 01:55:20.789 UTC"
## [5] "2015-12-04 02:06:13.294 UTC"
## [6] "2015-12-04 02:07:52.295 UTC"
## [7] "2015-12-04 02:13:20.798 UTC"
## [8] "2015-12-04 02:16:43.299 UTC"
## [9] "2015-12-04 02:46:16.313 UTC"
## [10] "2015-12-04 02:46:25.313 UTC"
## [11] "2015-12-04 04:31:07.364 UTC"
## [12] "2015-12-04 04:31:07.364 UTC"
## [13] "2015-12-04 08:53:05.992 UTC"
## [14] "2015-12-04 08:53:28.492 UTC"
## [15] "2015-12-04 10:13:30.032 UTC"
## [16] "2015-12-04 10:14:15.032 UTC"
## [17] "2015-12-04 10:21:18.036 UTC"
## [18] "2015-12-04 10:41:42.046 UTC"

## [19] "2015-12-04 10:45:18.047 UTC"
## [20] "2015-12-04 10:45:49.548 UTC"
## [21] "2015-12-04 10:57:04.553 UTC"
## [22] "2015-12-04 11:03:00.056 UTC"
## [23] "2015-12-04 11:05:55.558 UTC"
## [24] "2015-12-04 11:12:31.561 UTC"
## [25] "2015-12-04 23:51:12.509 UTC"
## [26] "2015-12-04 23:51:44.009 UTC"
```