PROJECT REPORT

Asteroid Data Classification and comparison using Raw and principal data

TECHNOLOGY & SYSTEMS

By

ASHRYA AGRAWAL

09.08.2019
2nd year Computer Science undergraduate
BITS Pilani - Pilani Campus, Rajasthan
Intern Batch : 2
Email : ashryaagr@gmail.com

Acknowledgement

Data-set: All the data is from the (http://neo.jpl.nasa.gov/). This API is maintained by SpaceRocks Team: David Greenfield, Arezu Sarvestani, Jason English and Peter Baunach. The problem statement is taken from kaggle. I would also thank AI-Techsystems for continuously monitoring the work and providing the support.

Abstract

Asteroids can be classified depending on whether they are hazardous or not. Machine Learning can be used to do this. ML models like SVM can classify the asteroids to a significant level of accuracy.

There is not much difference if we use raw data instead of principal components. Accuracy changes by less than 1 %. The training time on raw data is more than when we train on principal components .

Introduction

Asteroids can be a potential threat for life on Earth. Thus it is important to detect the asteroids that can potentially harm Earth, in order to take timely measures. As there are a large number of asteroids in space, it is essential that the process of finding (or classifying as) hazardous asteroids be done by computers based on the data from satellites. Machine Learning can be used for this.

This project uses SVM to classify the asteroids .

Description

SVM classifier of sklearn is used to perform classifier on both raw data and principal components.Then, before using SVM dates are converted to a number of days (relative to a fixed day) which is compatible with sklearn SVC. The strings like “Equinox”, “Orbiting Body” have been converted to a dictionary and corresponding keys are used.

The features that don’t help in classification but instead reduce the performance are dropped.

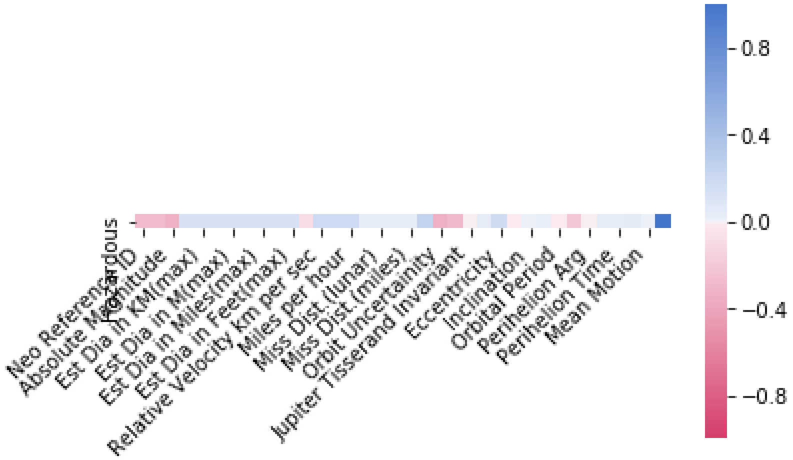
After this MinMaxScaler is used to scale the data. This scaling of data is necessary to avoid exceptionally bad results. Heat map from seaborn is used to visualise the correlation of the features.
Note : Correlation is a measure of only linear relationship. It can not be used to conclude if output does not depend on a particular feature.

The target i.e. “is_hazardous” is one hot encoded to feed into the model. Sklearn function train_test_split is used to create training and test dataset. Linear classifier is used together with OnevsRestClassifier.
After this we get accuracy and hyperplanes. Matplotlib.pyplot is used to visualise hyperplane and the dataset.

The above process is done both, while working on principal components and raw data.
But while working with raw data one additional step of loading data from various json files is also done.

Input Analysis

Heat map for correlation between features w.r.t. the feature “hazardous” :



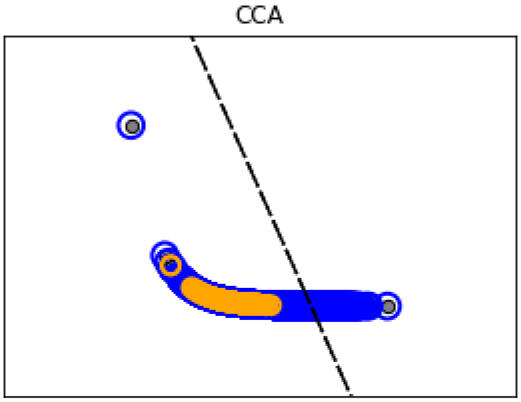
Observations

The accuracy and training time :

Using principal components of data :

Accuracy : 94.88%

Training time : 582ms

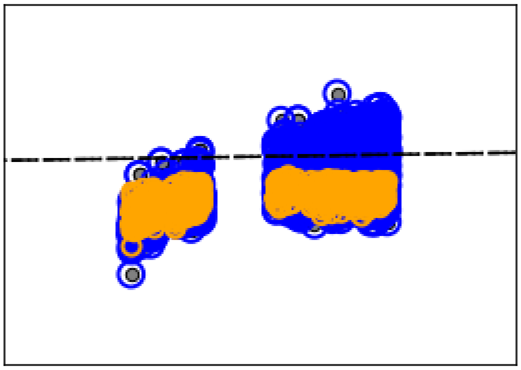


Using raw data :

Accuracy : 94.67%

Training time : 717ms

CCA



Conclusion

There is not much difference in accuracy in the two cases. The training time is slightly higher when we use raw data.

Scope in future

Machine Learning can definitely help in speeding up the process of identification of hazardous asteroids and space material in future. These objects on being detected and classified as hazardous, can be destroyed or their path can also be changed to avoid destruction on earth.