NASA SPACE APP CHALLENGE

NEAR EARTH OBJECTS

MACHINE LEARNING

ASTEROID FOREST

SKOPJE, APRIL 2016

With the ongoing problem with detection and characterization of asteroids, who could someday hit us, we decided to take on this challenge to help the scientists who work at NASA and the other astronomical societies around the world. Our idea is to use Machine Learning for the characterization of asteroids who could potentially be harmful.

Firstly, we obtained the data that the Minor Planet Center had posted on their official website. Specifically we downloaded the data set for “Orbits for Near Earth Asteroids (NEAs)” and the data set for “Orbits for Potentially Hazardous Asteroids (PHAs)”. Our idea was to compare the data sets and find which of the NEAs were PHAs, so we can add a new attribute (Hazardous) to this data set, who had two nominal values (T=true, F=false). Later we would use this modified data to train our model.

The dataset of NEAs was consisted of 14000 instances and all of the instances had 23 attributes. Initially we removed that attributes that weren’t described, so after the first preprocessing, the data had 18 attributes. The data set had many negative (Hazardous=F) examples, so in order to give the attribute a Uniform distribution, we removed random negative examples from the NEA set, so we ended up with 9941 instances, where half were positive examples and the other half negative. The next step in the preprocessing process was to find what attributes had valuable information. The attributes were:

* Des’n – number id of the instance
* H – Absolute visual Magnitude
* G – Slope parameter
* Epoch – The epoch of osculation of the orbital elements
* M – mean anomaly at the epoch
* Perihelion – The J2000.0 argument of perihelion (in degrees)
* Node - The J2000.0 longitude of the ascending node (in degrees)
* Inclination – The J2000.0 inclination (in degrees)
* e – Orbital eccentricity
* n – Mean daily motion (in degrees/day)
* a – Semi-major axis (in AU)
* Reference – where info for the asteroid can be found
* #Obs – number of observations
* #Opp – number of oppositions
* Arc – arc length (days)
* rms – root mean square
* Perts – Precise indicator of pertrubers
* Computer – on what computer was this data calculated

We wanted to have information that somehow divided certain asteroids in family of asteroids who will have similar attribute values. Therefore information from the attributes Computer, Arc Perts, Reference, #Obs, #Opp and Des’n were not needed and didn’t give us information about the asteroid so they were removed from the training set. In the end we ended up with 13 attributes. Because the continuous attributes had values who weren’t in a small range, we decided to normalize all of the continuous attributes with a z-score normalization. All of the values from a given attribute were subtracted with the mean average of the attribute, and then divided by the standard deviation of the attribute. That gave us values that would be in a small range, so of the attributes would have the same weight in the decision process. We also tried to find correlation between the attributes but there wasn’t anything significant.

The next step was finding the best suited model for classification. We used the application Weka in this process. We tried out Naïve Bayes Classification (as a benchmark classification so we can measure the performance of other classifier based on the simplest classification model). After that we used Multilayer Perceptron (Neural Networks), RBFNetwork, VotedPerceptron, a C4.5 tree classification and Random Forest classification. We tested all models using Cross-Validation with k=10 folds. From the results it was obvious that Random Forest model had the best evaluation results. The next better accuracy was 16% less than the Random Forest model. These were the results we got from the model evaluation.

# Random Forest

We used 100 trees, and every tree considered 4 random attributes

|  |  |  |
| --- | --- | --- |
| Correctly Classified Instances | 78.7159 % | 7822 |
| Incorrectly Classified Instances | 21.2841 % | 2115 |
| Kappa statistic | 0.5744 |  |
| Mean absolute error | 0.315 |  |
| Root mean squared error | 0.3797 |  |
| Relative absolute error | 63.0068 % |  |
| Root relative squared error | 75.9479 % |  |
| Total Number of Instances | 9937 | |

### Detailed Accuracy by Class

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
| 0.756 | 0.182 | 0.806 | 0.756 | 0.781 | 0.883 | T |
| 0.818 | 0.244 | 0.77 | 0.818 | 0.793 | 0.883 | F |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Weighted Avg. | 0.787 | 0.213 | 0.788 | 0.787 | 0.787 | 0.883 |

### Confusion Matrix

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
| a b <-- classified as | | |
| 3763 | 1212 | a = T |
| 903 | 4059 | b = F |

From the results we can see that the model classified 78.7159 % of the test examples correctly and that the Root mean squared error is 0.3797. The confusion matrix shows that a large number of hazardous and not hazardous asteroids are classified correctly. With the feature selection made earlier and quest to find the best model we also wanted to minimize the error. Errors that occurred when the model was classifying hazardous asteroids as not hazardous can be really dangerous. This model with the Random Forest classifier gave us best results and the fact that it had 100 trees with 4 random attributes made this model very good in generalization (so it wasn’t overfitting the training set).

The accuracy we achieved is satisfying and the model can still be very useful because it is based on decision trees and classification is really fast. It can be used for classifying the asteroids the same moment we get the data which can be very useful for detecting danger and provide us with more time to solve the threat.

We are glad that we were working on this particular problem because it affects the whole planet. The challenge helped us learn more about Machine Learning, dealing with big data, but it also helped us learn space, asteroids and the dangers the surround our planet Earth.