International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Hazardous Asteroid Classification through Various Machine Learning Techniques

Anish Si

Student, Master of Computer Application, School of Information and Technology, Vellore Institute of Technology, Tamil Nadu, India

---

Abstract - *In this paper, I tried to find a new model to identify the hazardous asteroids. As far as, asteroids are concerned, there are many asteroids called near-earth asteroids, but all are not hazardous. So, our target in this paper is to identify those hazardous asteroids and classify them with non-hazardous types. For this, I choose many machine learning models. I trained those various model with the data features and later I compare those results to find the most accurate model which gives the most accurate classification. In my prediction random forest and xgbclassifier give most accurate prediction.*

Keywords: **Hazardous, Non-hazardous, Machine learning, Random forest, Xgbclassifier**

1. INTRODUCTION:

Asteroids are one of the most discussed material of the space. It is one of the most important material to find the characteristics of the solar systems. So it is very much valuable to the scientist. It is divided into various classes. Some of them are MCA, MBA, TRO etc. But in this division all are not the hazardous, only the asteroids move near the earth are potentially hazardous. They are of four types. But all are not hazardous also. Those are known as Atens, Atiras, Amors, Apollos. They are divided by various features such as semi major axis, eccentricity etc. So the hazardous and non-hazardous quality is also depending on some attributes such as its diameter, relative velocity, magnitude, inclination etc.

It has been seen that now-a-days, machine learning is the one of the most important technique for predicting or classifying the dataset. So here, I use some of the machine learning model and later compare those results to show which one of them is the more accurate for classifying the asteroids as hazardous or non-hazardous.

2. LITERATURE REVIEW:

In this paper[1],the authors tried to make some predictions over the combinations of orbital parameters for yet undiscovered and potentially hazardous asteroids by machine learning techniques. For this reason, they have used the Support Vector Machine algorithm with the kernel RBF. By this approach, the boundaries of the potentially hazardous asteroid groups in 2-D and 3-D can be easily understood. By this algorithm, they have achieved the accuracy over 90%.

In this paper[2], authors has provided classification of asteroids, which are observed by VISTA-VHS survey. They have used some statistical methods to classify the 18,265 asteroids. They used a probabilistic method, KNN method and a statistical method to classify those. Later, they compared the algorithms' accuracy. They have classified the asteroids into V, S and A types. They did it on the 18265 asteroids and test set consists of over 6400 asteroids.

In this Paper[3], they have classified the asteroids which are near to the earth and can easily pass through the orbital of Mars. For this purpose they, have used a perceptron-type network. And their output class are three, they are Apollo, Amor and Aten. They have also found that semi-major axis and focal distance are the two important major features, for which the asteroids can be separable linearly.

In this paper[4], they identified the asteroid family by the help of the supervised hierarchical clustering method. For this method, they have found the distance between an asteroid from any reference objects. Later they compared the results with traditional hierarchical clustering method and found that their result gives 89.5% accuracy, which is better. By this approach, they found 6 new families and 13 clumps.

In this paper[5], they used KNN, gradient boosting, decision trees, logistic regression to classify the family of the asteroids. They have identified main belt asteroids, which are three body resonant. They have identifies 404 new asteroids. They found that gradient boosting methodgives the accuracy of 99.97% for identification purpose which is maximum among all the methods they have applied.

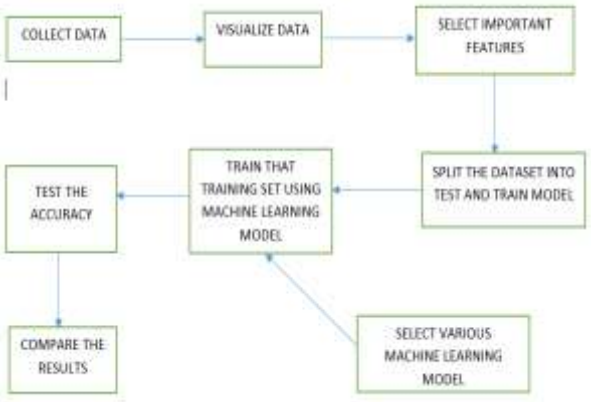
In this paper[6], they have also classified the asteroids according to their family such as Apollo, Amor, Aten etc. They have done it only over the asteroids, who pass the mars orbit. They can be easily separable by the semi major axis and the focal distances. So, a perceptron type artificial

neural network is enough to classify those three types of asteroids.

In this paper[7], the author has used random forest techniques to classify the asteroids. They have done it on the 48642 asteroids. They have classified those asteroids among 8 classes such as C, X, S, B, D, K, L and V.This division are done according to their SDSS magnitudes at the wavebands of g, r, I and z. They also reached the higher accuracy compare to the many proposed methods.

In this paper[8], they used supervised learning algorithm to detect asteroid from a large dataset. For this case they have used vetted NEOWISE dataset (E. L. Wright et. Al,2010). According to them, the metrics they have used can be easily done as it can be easily associated with the extracted sources. They also used the python SKLEARN package. After doing this also gave the report on the reliability, feature set selection, and also the suitability of the various machine learning algorithms. At the end they also compares the results.

3. METHODOLOGY:



3.1 Data collection:

This data is taken from the Kaggle. This is the data consists of 4688 rows and 40 columns. All the data is from the (http://neo.jpl.nasa.gov/).

3.2 Data Visualization:

Here, for data visualization purpose we mainly use heat map, which is done by the help of the pearson correlation coefficient. From this we can easily identify the highly correlated columns and take necessary steps to remove the dependency.

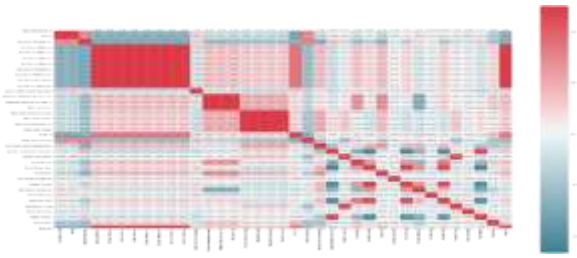


Fig 1. Heat map plot

After checking of duplicate values and missing values, we found that the count of hazardous and non-hazardous asteroids. There are only 755 asteroids are hazardous while remaining 3932 are non-hazardous.

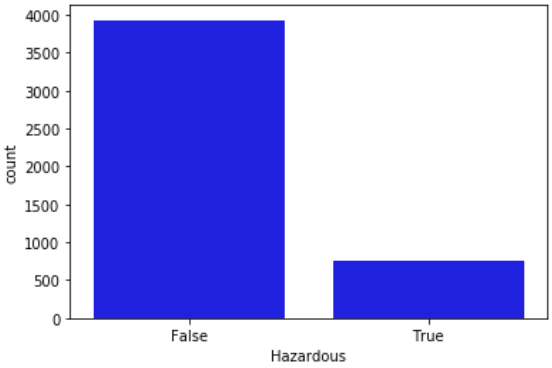


Fig 2. Count plot

3.3 Features selection:

For this classification I selected 15 features. They are absolute magnitude, minimum orbit intersection, Ascending node longitude, orbit uncertainty, perihelion time, inclination, semi major axis, Anomaly, perihelion arguments, perihelion time, relative velocity, perihelion distance, eccentricity, aphelion distance and Jupiter Tisserand Invariant. These are the important features. Here the features are ordered from high to less important. Some of the less important features are also depended on the other highly important and independent features such as Jupiter Tisserand Invariant which is dependent on semi-major axis, orbital eccentricity and inclination.

3.4 Split dataset:

Then, the dataset is divided into test and train set. Test set has 938 data while training set comprise of 3749 data. Now the whole dataset is suitable for using the machine learning techniques. I am splitting this by taking the random state value as 10.

3.5 Techniques and their accuracy:

For implementing this, I use python. Because it has many libraries such as numpy, scikit-learn, pandas etc. Here I use matplotlib, seaborn, os and xgboost with the above mentioned three libraries.

Here I choose eight machine learning techniques. They are logistic regression, support vector machine, decision tree, K nearest neighbor, random forest, naïve bayes, adaboost and xgboost method. Of them random forest and xgboost methods are giving the highest accuracy of 100% while, Naïve Bayes method gives the lowest accuracy of 80.70%.

	accuracy	precision Not hazardous	precision hazardous	recall Not hazardous	recall hazardous	training time
Log reg	0.829424	0.832439	0.426571	0.994865	0.019868	0.274791
Log reg gridsearch	0.827292	0.842387	0.459459	0.974326	0.109818	16.614259
SVC	0.831557	0.831577	1.000000	1.000000	0.066289	2.804657
knn	0.830490	0.830490	0.000000	1.000000	0.000000	0.060777
dec_tree	0.987908	1.000000	0.967578	0.987433	1.000000	0.159087
Rand Forest	1.000000	1.000000	1.000000	1.000000	1.000000	0.174722
naive_bayes	0.807038	0.833795	0.238095	0.958922	0.062893	0.016789
Adaboost	0.828358	0.832256	0.375000	0.993582	0.019868	14.577214
xgb	1.000000	1.000000	1.000000	1.000000	1.000000	2.400512

Fig 3. Final comparison

From this table we can see that random forest (with the number of tree is 15) and xgbclassifier can classify this hazardous and non-hazardous asteroid most accurately, but between them random forest takes minimum training time. Random forest takes the time 0.174722 ms, while XGBclassifier takes 2.400512 ms. So, random forest with n_estimators of 15 should be the better approach to classify the asteroids.

4. CONCLUSION:

In this paper, I have done this classification of hazardous and non-hazardous through various machine learning techniques. After using many methods, it has been seen that random forest with tree number 15 is the most optimal model according to accuracy and training time. In my opinion, this paper will help to identify the newly discovered near earth asteroids if that is hazardous or not.

5. REFERENCES:

[1] Prediction of Orbital Parameters for Undiscovered Potentially Hazardous Asteroids Using Machine Learning, Vadym Pasko
[2] Taxonomic classification of asteroids based on the Movis near Infrared-Color, M. Popescu
[3] Classification of Near Earth Asteroids by Artificial Neural Network, Ihara Csllik
[4] Machine-learning identification of asteroid groups,V. Carruba , S. Aljbaae and A. Lucchini
[5] Identification of asteroid trapped inside three-body motion resonance: a machine learning technique, Sirnov, Markov
[6] Classifications of the near earth asteroids with artificial neural network, Zoltan Mako
[7] Spectral Classification of Asteroids by Random Forest, Huang Chao
[8] Machine learning and next generation asteroid surveys,Nunget, Carry