International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

A HYBRID METHOD TO ENHANCE THE PREDICTION OF HAZARDOUS ASTEROIDS USING XGBOOST CLASSIFIER WITH XGBCLASSIFER BASED FEATURE SELECTION METHOD

Nerella Tarun Reddy[1], G Jagapathi Reddy[2], A Shashikanth[3], Dr. Vijay Anand[4]

[1]Student,Computer Science and Engineering, JBIET, Telangana, India.
[2]Student, Computer Science and Engineering, JBIET, Telangana, India.
[3]Student, Computer Science and Engineering, JBIET, Telangana, India.
[4] Assistant Professor, Computer Science and Engineering, JBIET, Telangana, India.

Abstract -*A multitude of variables influence an asteroid's route, all of which can cause the asteroid's path to alter at any time. It is nearly impossible to predict asteroid collisions and devisemitigation strategieswithout the aidof computational power. Through automated solutions like NASA's Goldstone Deep Space Communications and Spacewatch's Moving Object Detection Program an enormous amount of data is collected every minute and it's difficult to process it all in real time with a high degree of accuracy. According to preliminary studies, the machine learning algorithms produce good results. In this work, a novel method is proposed to enhance the performance of XGBOOST algorithm with XGBclassifier based feature selection algorithm. Along with, this research evaluates a variety of machine learning techniques as well as feature selection algorithms to determine the most effective way for forecasting hazardous asteroids.*

***KeyWords*: XGBOOST, SVM, RANDOM FOREST, XGBCLASSIFIER, MUTUAL INFORMATION**

1. INTRODUCTION

Asteroids are rocky celestial objects, they are the primeval remains of early formation of our Solar System, which occurred around 4.5 billion years ago. Majority of the asteroids are present in the space between the orbits of Mars and Jupiter. Earth has been struck many times by asteroids whose orbits hauled them into the central solar system , aggregately known as Near Earth Asteroids (NEAs), still a threat to Earth today. Just to depend on our luck would be complacent after having the intellectuality that humans have developed so far. The impact of an asteroid could wipe out the human race, like it happened to the dinosaurs. Earth is impregnable when it comes to the case of non-hazardous asteroids.

The asteroids which have an absolute magnitude of 22.0 or less or have diameter greater than about 140 metres and an Earth Minimum Orbit Intersection Distance of 0.05 au (7,479,894 Kilometres) or less are considered as Potentially Hazardous Asteroids (PHA's)[1]i.e it's orbit has a chance of collision with Earth and has an adequate scope to cause substantial and prolonged damage during the occurrence of impact.

About 30% of the examined orbits have MOID less than 0.05 au [2] . In order to forestall a grand-scale turmoil by PHA's we need years of devising. Accordingly, we necessitate an adequate admonition period to assemble a plan of action [3].There are numerous automated solutions to detect the asteroids, such as Spacewatch's Moving Object Detection Program which uses image processing and NASA's Goldstone Deep Space Communications uses radar, to divulge particulars such as its shape, size and whether the asteroid is a single object or a system of one small object revolving around a larger one. Also, in near future, we have impending technologies like The Large Synoptic Survey Telescope which could enhance and calibrate the search for NEO's with improved accuracy.[4]. Through such methods, an enormous amount of data is collected every minute and it's difficult to process it all in real time with a high degree of accuracy. To face this challenge, we apply machine learning and feature selection techniques to scrutinize the data. Maneuvering machine learning algorithms like Support Vector Machines, XGBoost and Random Forest have shown promising results in other fields, and might be able to deal with enormous amounts of data that has to be scrutinized.

2. PROPOSED METHOD

In the proposed work, the performance of XGBOOST classifier is improved by selecting the informative features using XGBclassifier based feature selection method. The block diagram of the proposed method is given in the Fig.1

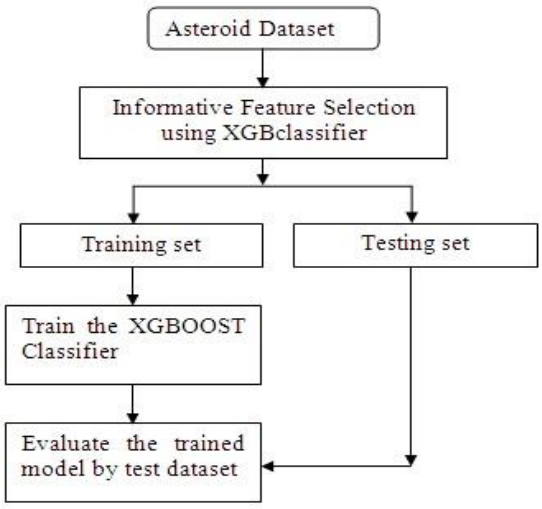

Fig -1: Block diagram of Proposed method

3. METHOD SECTION

The dataset employed from kaggle embodies 40 features. To record these observations, NASA has used various Ground-based telescopes and NEOWISE (Near-Earth Object Wide-field Infrared Survey Explorer) spacecraft . This data is managed by the Jet Propulsion Laboratory (JPL) and the operations, data processing and archiving takes place at Infrared Processing and Analysis Center (IPAC) at California Institute of Technology (CALTECH).

Supervised learning uses labeled datasets to train algorithms that classify data or predict outcomes accurately. Supervised learning can be split into classification and regression problems. We use the classification algorithms like Random Forest, SVM etc. to recognize the features within the dataset in order to label the data. Regression algorithms like mutual information are used to know the relationship between two or more dependent and independent variables. Since we already have the data of a few hazardous and non hazardous asteroids in our solar system, we can use the features of those asteroids to train our algorithm to further predict if a newly detected asteroid is hazardous to earth.

The dataset includes 4687 observations. We split the dataset into training and testing sets in the ideal ratio of 8:2.The constraints for an asteroid to be hazardous or not are studied using training data.

3.1 XGB

It is nearly impossible for the traditional algorithms to deal with the immense amount of data involved with observational astronomy. The XGBoost model comes with preprogrammed multithreading parallel computing, generalisation ability, tree structure, can manage missing values and such features which coalesce it into a much faster and suitable method to deal with our data. XG Boost is extensively used in many fields to accomplish some data challenges, and also it is a highly operative scalable ML system for tree boosting. XGB is among the efficacious methods due to its features such as easy parallelism and high prediction accuracy[5].

3.2 SVM

SVM is a memory efficient algorithm as it uses a subset of training points in support vectors. It is also proven to be effective in high dimensional spaces.SVMs reduce most machine learning problems to optimization problems and optimization lies at the heart of SVMs [6]. During the training phase of the algorithm, the SVM reads in a set of data points whose categories are known. The goal of the algorithm is to be able to separate the data into two groups when the dataset is plotted in space. To do this, it attempts to draw a hyperplane between the two clusters of points[7].

3.3 RANDOM FOREST

The Random Forest is appropriate for high dimensional data modeling because it can handle missing values and can handle continuous, categorical and binary data. Accurate predictions and better generalizations are achieved due to utilization of ensemble strategies and random sampling[8]. Random Forest uses Decision Trees as base classifier. This ensemble learning method is used for classification and regression of data. Random forest is an ensemble method which generates accurate results but, on the other hand it is a time consuming method too as compared with the other techniques [9].Accurate classification of the dataset can be done by generating a forest of decision trees (Random Forest algorithm) via supervised learning [10].

4. FEATURE SELECTION

Only a few traits out of forty play a significant role in determining the asteroid's path. As a result, we apply feature selection methods to reduce the number of elements in order to improve prediction precision. To overcome the effects of atmospheric seeing on the detection of NEO's using ground based telescopes, the concept of Rayleigh resolution is used, similarly we use feature selection algorithms to refine the data that is being analysed[11]. The goal of feature selection methods is to prevent the use of a model that contains irrelevant features or components.

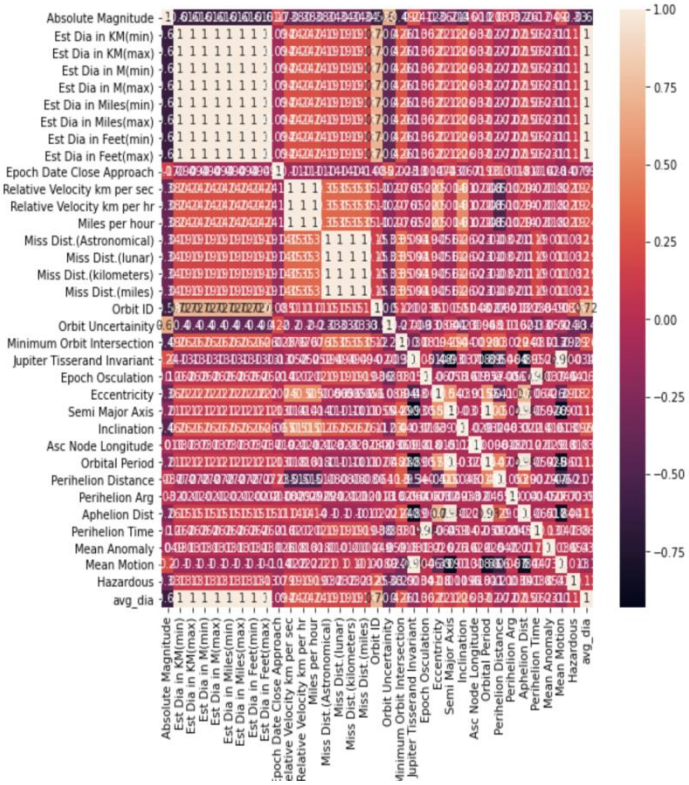

Fig -2: Heat Map

4.1 MUTUAL INFORMATION

Mutual Information The mutual information is a general measure of the dependence between two random variables. It expresses the quantity of information one has obtained on X by observing Y . The mutual information is always greater than or equal to zero, with equality iff X and Y are independent. It is lower than the entropy of either variable, and equality only occurs if one variable is a deterministic function of the other. The higher the mutual information, the stronger the dependency between X and Y[12].

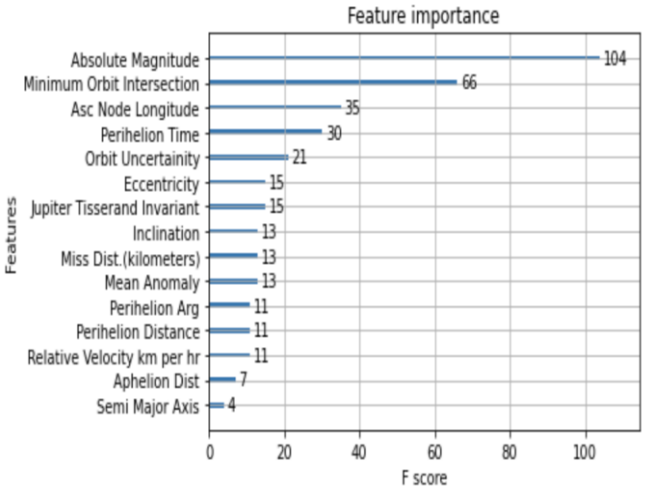

Chart -1: Feature Importance

5. RESULTS

We analyze the test datasets after training the XGBOOST, SVM, Random Forest models on the whole training datasets separately. The accuracy of XGB with XGBCLASSIFIER feature selection was 99.74 percent. With MI feature selection, XGB was 99.68 percent accurate. Random forest with mutual information gave an accuracy of 99.67%. Random Forest with XGBCLASSIFIER feature selection produced an accuracy 99.61%. The accuracy of SVM with XGBCLASSIFIER feature selection is 88.82%. SVM with mutual information feature selection is 83.47% accurate.

Table-1: Comparison of proposed method with other algorithms

Algorithms	Accuracy (%)
SVM	73.23
Random Forest	74.56
XGBOOST	84.23
SVM + MI	83.47
Random Forest + MI	92.51
XGBOOST + MI	99.61
SVM + XGBClassifier	88.82
Random Forest + XGBClassifier	93.21
XGBOOST + XGBClassifier	99.74

6. CONCLUSION

This study analyses feature selection strategies in dimensionality reduction at different levels, from using all features to removing a large number of them. The greatest results came from XGBOOST with XGBCLASSIFIER, followed by XGBOOST with MI and Random Forest with MI. When compared to other algorithms, SVM performed badly, but when combined with XGBCLASSIFIER for feature selection, It showed a considerable improvement in performance.

REFERENCES

[1] Robert Jedicke, Alessandro Morbidelli, Timothy Spahr, Jean-Marc Petit, William F. Bottke, Earth and space-based NEO survey simulations: prospects for achieving the spaceguard goal, Icarus, Volume 161, Issue 1, 2003, Pages 17-33, ISSN 0019-1035, https://doi.org/10.1016/S0019-1035(02)00026-X.

[2] C. de la Fuente Marcos, R. de la Fuente Marcos, Asteroid 2013 ND_{15}: Trojan companion to Venus, PHA to the Earth, Monthly Notices of the Royal Astronomical Society, Volume 439, Issue 3, 11 April 2014, Pages 2970–2977, https://doi.org/10.1093/mnras/stu152

[3] B. D. Seery et al., "Near Earth object mitigation studies," 2016 IEEE Aerospace Conference, 2016, pp. 1-12, doi: 10.1109/AERO.2016.7500546.

[4] S. R. Chesley and P. Vereš, "The Large Synoptic Survey Telescope: Projected near-Earth object discovery performance," 2016 IEEE Aerospace Conference, 2016, pp. 1-8, doi: 10.1109/AERO.2016.7500539.

[5] M. Chen, Q. Liu, S. Chen, Y. Liu, C. Zhang and R. Liu, "XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System," in IEEE Access, vol. 7, pp. 13149-13158, 2019, doi: 10.1109/ACCESS.2019.2893448.

[6] Tian, Yingjie & Shi, Yong & Liu, Xiaohui. (2012). Recent advances on support vector machines research. Technological and Economic Development of Economy. 18. 10.3846/20294913.2012.661205.

[7] M. Freed and J. Lee, "Application of Support Vector Machines to the Classification of Galaxy Morphologies," 2013 International Conference on Computational and Information Sciences, 2013, pp. 322-325, doi: 10.1109/ICCIS.2013.92.

[8] Ali, Jehad& Khan, Rehanullah& Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9

[9] Goel, Eesha and Er. Abhilasha. “Random Forest: A Review.” (2017).

[10] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[11] O'Dell, Anthony & Cain, Stephen. (2009). Investigating the effects of atmospheric seeing on the detection of near earth orbiting asteroids. IEEE Aerospace Conference Proceedings. 1-9. 10.1109/AERO.2009.4839455.

[12] Batina, L., Gierlichs, B., Prouff, E. et al. Mutual Information Analysis: a Comprehensive Study. J Cryptol 24, 269–291 (2011). https://doi.org/10.1007/s00145-010-9084-8