Meteor Shower Scale Prediction Using Random Forest Classification

Bozhen Peng*, Dingyi Sun, Qifei Cui, Chun Yin Yip
Chengdu Shude High School, Chengdu, China
boby.pengr@outlook.com

Abstract—With the development of science and technology, the aerospace industry also has been developing rapidly. At the same time, meteor showers from the universe become one of the severest threats to the increasing numbers of aero crafts in the outer space. With the advancement of computer science, we have been able to employ big data to analyze and predict the scale of meteor showers, so as to reduce the possibility of meteor objects hitting spacecraft and provide a strong guarantee for aerospace security. The Random Forest Classification, a kind of machine learning algorithm, is used to analyze meteor showers that once visited the earth to explore the principle of meteoroid cycles and predict the scale of meteor showers that will enter the Earth. The confusion matrix and accuracy rate will be used to evaluate the performance of the machine learning model.

1. INTRODUCTION
In recent years, the development of China's aerospace industry has entered a new era. Since its establishment in 1956, it has reached a considerable scale and level after several important periods. [1] In outer space, the number of satellites is increasing, and the orbit is becoming denser.

The problem therefore occurred: the aircraft may have a running orbit overlapping the meteoroid falling orbit. That is, an aircraft may be hit by the meteoroid, posing a major security threat to personnel and equipment. Though the spacecraft do have protective measures, the possibility of the damage of solar panels, the most vulnerable part of satellites, would be increasingly high when dense meteoroids enter the Earth's atmosphere. In general, satellites change the direction of the solar panels to minimize the chance of damage. However, during the 1993 Perseid meteor shower, the European Space Agency's "Olympus 1" satellite was exposed to a large number of meteoroids. The airborne gyroscope was crushed by a meteoroid [2], and eventually the satellite lost control and could not adjust its posture to avoid the meteor attack.

The Perseid meteor shower seemed to have caused the "Olympus 1" satellite to lose control due to too fast spin. The European Space Agency had successfully restored the position of the satellite, but it consumed most of the satellite's fuel. The remaining fuel of the satellite is only enough to send it to the Graveyard orbit [3]. The European Aerospace Agency’s “five-year plan” had to end early.

This work will use machine learning to study and analyze the scale of meteor showers, and thus better help aerospace professionals to design and change spacecraft orbits, reduce the possibility of aerospace disasters, and contribute to the development of aerospace industry.
2. METHODS

2.1 Data pre-processing
We extracted data from Visual Meteor Database of International Meteor Organization, IMO. It contains meteor data recorded with standardized observations during the 39 years from 1980 to 2018, all of which were uploaded by IMO members. The dataset is divided by year, and an annual data is divided into two separate files, Session and Rate. The features of Session include Session ID, Start Data, User ID, First Name, Last Name, City, Country, Latitude, Longitude, and Elevation. The features of Rate are Rate ID, User ID, Obs Session ID, Start Date, End Data, Ra, Decl, Teff, F, Lm, Shower, Method, and Number.

For model development, we merged the annual files of Rate and Session into two files, respectively. Then, using Session ID in Session and Obs Session ID in Rate as the common labels, we merged the Rate file and Session file. Occurring Month is extracted from Start Date. Three features, including Longitude, Latitude, and Elevation, are merged into a single feature, Position. Other irrelevant features were removed.

This work examines how a given feature vector can be used to predict the scale of the meteor shower described by the vector. After pre-processing, a total of 550,728 valid data were obtained for training machine learning models, which were divided into independent variables and dependent variables. The independent variables are Occurring Month (Occuring Month), Right ascension (Ra), Declination (Decl), Shower Type (Shower), Observing Location (Position), and the dependent variable is the number of meteors (Number).

The features of non-integer variables or floating-point variables in the dataset—Shower and Position—are normalized to floating-point variables in the range [-1, 1], making the features comparable while speeding up the fitting of machine learning models[4]. This work defines the meteor shower scale (Class) based on the number of meteors (Number)[5]. If the number of meteors is in the range [0, 5], the meteor shower is defined as Small-sized, with the Class value of 1; if the number of meteors is within the range of (5, 10], the meteor shower is Medium-sized, with the Class value of 2; if the number of meteors is within the range of (10, ∞], the scale of the meteor shower is defined as Large-sized, with the Class value of 3. The final output of this work will assign 1, 2, or 3 to the value of meteor shower scale (Class). Then, 495 variables in the Shower and 12 variables in the Occurring Month are encoded into dummy variables [6] to avoid over-fitting and boost training efficiency.

| Type     | Name          | Description                               |
|----------|---------------|-------------------------------------------|
| Independent | Occurring Month | The month in which a meteor shower occurred |
|          | Ra            | Right ascension                           |
|          | Decl          | Declination                               |
|          | Shower        | Shower Type                               |
|          | Position      | Location where a meteor shower occurred, zipped from Longitude, Latitude, and Elevation |
| Dependent | Class         | The scale of a meteor shower               |

2.2 Random Forest Classification
This paper employed the Random Forest Classification[7] to predict the scale of meteors based on feature vectors. The Random Forest Classification is a kind of Ensemble Learning Algorithm, which is based on the Decision Tree Classification and has an overall improvement in efficiency and accuracy compared to the latter. The Decision Tree is a tree-like structure built on the basis of decision-making and contains a large number of branches as nodes to divide the data.
We used the ID3 (Iterative Dichotomiser 3) algorithm [8] to build decision trees in the random forest. In the ID3 algorithm, the increase in information entropy caused by the arrangement will determine where each feature classification appears in the decision tree. The ID3 algorithm selects the arrangement that minimizes the "purity" of the data to build a decision tree. In general, attributes that divide information into as many subsets as possible have a higher "purity" and should be placed closer to the root node of the decision tree. The important physical quantity that quantifies the purity of information is information entropy. An effective arrangement will make the information gain as large as possible: the information uncertainty is reduced as much as possible, thereby reducing the overall information entropy and improving the "purity".

The entropy of the random variable $X$ can be expressed by the formula:

$$
Ent(X) = - \sum_{k \in V(X)} p_k \log_2 p_k
$$

(1)

where $X$ is a random variable, $V(X)$ represents a set of all possible values of the variable $X$, and $p_k$ represents the probability that $X$ takes a value $k$.

Thus, the empirical entropy of the dataset is expressed as

$$
Ent(D) = - \sum_{k \in V(D)} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|}
$$

(2)

where $D$ is the set of samples in dataset $D$, $V(D)$ represents the set of all values of feature classification of samples in the dataset, $|C_k|$ represents the number of samples whose feature is classified as $k$, and $|D|$ is the total number of samples.

Next, we need to calculate the information gain brought by each feature classification. Because the information gain reflects the difference in information uncertainty before and after classification, the greater information gain makes the divided information more representative. Before each node division of the decision tree, the algorithm calculates the information gain brought by each feature classification and selects the feature classification with the largest information gain to divide, so as to maximize the ability of the model to distinguish samples. The algorithm calculates the information gain by the following formula:

$$
Gain(D,A) = Ent(D) - Ent(D|A)
$$

$$
= Ent(D) - \sum_{v \in V(A)} \frac{|D_v|}{|D|} Ent(D_v)
$$

(3)

Where $Ent(D|A)$ represents the conditional entropy of the sample under the condition of feature classification $A$, $D$ is the set of all samples in data set $D$, and $A$ is a feature used to divide the data set, $V(A)$ Representing a set of all values of feature class $A$, $v$ is one of the values of feature classification $A$, and $D_v$ represents a set of samples with feature classification $A$ whose value is $v$ in dataset $D$.

A large number of different decision trees together constitute a random forest. The random forest classification is based on modified Bagging. First, the algorithm will use Bootstrap sampling in the N samples of the total training set, i.e. randomly extracting n (n < N) training samples with replacement as a training set for one tree in the random forest; bootstrap sampling is also used for feature classification, where m (m < M) feature subsets are randomly chosen from M feature classifications with replacement. Subsequently, each decision tree is allowed to grow unrestrictedly and split as much as possible without any pruning. For a newly input sample set, T decision trees in the random forest will generate T classification results and do Majority Voting, and the classification result with the most votes will be final output of the forest.
2.3 Parameters
This prediction of the meteor shower scale is based on supervised learning, and this work will use the random forest algorithm to study the problem. Before the algorithm generates the machine learning model, the program divides the independent variables into training sets and test sets, randomly extracting 80% of the data from the data set as the training set and 20% of the data as the test set [9]. The training set will be used for random forest classification model fitting; the test set will not involve in the fitting of the random forest classification model and will be used as newly input data to verify the performance of the resulting model.

During the training process, the number of decision trees in the random forest will be set to multiple values (10, 30, 50, 70, 90) to predict the size of the meteor shower, and different confusion matrices [9] will be built for each to calculate the accuracy, and finally a value with the highest prediction accuracy will be chosen for the final model.

2.4 Performance Evaluation
Regarding the evaluation of the model, the following nine situations need to be considered:

- **Accurate**
  - \( f(X) = 1 \land y = 1 \),
  - \( f(X) = 2 \land y = 2 \),
  - \( f(X) = 3 \land y = 3 \),

- **Error**
  - \( f(x) = 2 \land y = 1 \),
  - \( f(x) = 3 \land y = 1 \),
  - \( f(x) = 1 \land y = 2 \),
  - \( f(x) = 3 \land y = 2 \),
  - \( f(x) = 1 \land y = 3 \),
  - \( f(x) = 2 \land y = 3 \),

Therefore, a confusion matrix can be established:
TABLE II. Result shown by confusion matrix

| Actualy | Confusion Matrix |
|---------|-----------------|
|         | 1               | 2          | 3           |
| 1       | Accurate₁       | Error₂₁    | Error₃₁     |
| 2       | Error₁₂        | Accurate₂  | Error₃₂     |
| 3       | Error₁₃        | Error₃₃    | Accurate₃   |

The Accuracy Rate of the model is shown:

\[
Accuracy\ Rate = \frac{\sum_{i} Accurate_{i}}{Total} \quad (4)
\]

Where \( i \) is the assigned value to a dependent variable (meteor shower scale), \( Accurate_{i} \) is the number of samples with accurate predictions and assigned \( i \), and \( Total \) is the total number of samples in the test set.

3. RESULT

The result is derived from the random forest model generated by the 50 decision trees and is represented by the confusion matrix.

TABLE III. Result of 50 decision trees in random forest shown by confusion matrix

| Confusion Matrix | Predicted f(X) |
|------------------|----------------|
|                  | 1              | 2              | 3              |
| Actualy          | 1              | 2              | 3              |
| 1                | 124986         | 4636           | 3293           |
| 2                | 9327           | 5627           | 4819           |
| 3                | 3975           | 3269           | 13420          |

Accuracy Rate = \( \frac{\sum_{i} Accurate_{i}}{Total} = \frac{144033}{173352} = 83.1\% \quad (5) \)

4. CONCLUSION

In this work we utilized random forest algorithm to predict the scale of the meteor shower. Under the random forest that limits the generation of 50 trees, the model has obtained a high prediction rate of 83.1\%, and the model accuracy can be improved by generating more decision trees. Predicting the size of the meteor shower can provide an early warning to the aerospace vehicles surrounding the outer atmosphere of the Earth's atmosphere, allowing them to actively avoid the outer space with high-density meteoroids entering and reducing the possibility of aerospace vehicles being out of control by the impact of meteoroids.

Meteor showers burn into dust in the atmosphere, providing additional condensation centers for cloud water vapor, increasing local cloud thickness and rainfall [10]. With the prediction of the scale of meteor showers, the study provides useful data for local meteorological conditions. This work will provide effective assistance for solar system research.

REFERENCES

[1] X. Zhang, “Status Quo of Chinese Aerospace and Its Prospect,” https://wenku.baidu.com/view/652988353186bceb18e8bbca.html, 2018/6/23
[2] S. Close, “Shooting Stars Can Shoot Down Satellites,” https://spectrum.ieee.org/aerospace/satellites/shooting-stars-can-shoot-down-satellites

[3] “Where does satellite go after being scrapped,” http://mini.eastday.com/mobile/170509184839799.html

[4] “Importance of Feature Scaling,” https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html

[5] K. Viswesh, “Machine Learning Analyses of Meteor Data,” WGN, The Journal of the IMO, https://www.cfrce.com/images/Viswesh_Paper1.pdf

[6] Katerina, “Working with categorical data,” https://medium.com/whats-your-data/working-with-categorical-data-c338122b9521

[7] L. Breiman and A. Cutler, “Random Forests,” https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#micro1

[8] J. R. Quinlan, “Induction of Decision Trees,” http://hunch.net/~coms-4771/quinlan.pdf

[9] Z. Zhou, “Machine Learning,” Beijing, Tsinghua University Press, 2016

[10] M. Yu, “Meteor shower observation and its research significance,” Journal of Beijing Normal University (Natural Science Edition), 2005, vol. 41