Exoplanet Classification with Data Mining

Israel dos Santos[a,b] [1], Adriana Valio[a,c], Nizam Omar[a,b] and A. H. F. Guimarães[c]

[a]Electrical Engineering and Computation Graduate Program (PPGEEC),
Mackenzie Presbyterian University, São Paulo, SP, Brazil
[b]Faculty of Computation and Informatics (FCI),
Mackenzie Presbyterian University, São Paulo, SP, Brazil
[c]Center for Radio Astronomy and Astrophysics Mackenzie, Engineering School,
Mackenzie Presbyterian University, São Paulo, SP, Brazil

Received on November 21, 2016 / accepted on December 30, 2016

Abstract

This article presents a case study on Data Mining in Astronomy to classify exoplanets and help the identification of terrestrial planets within the habitable zone of their host stars. To achieve this objective, Data Mining techniques were applied on the planets listed on the Exoplanet Orbit Database (EOD). These planets can be grouped into at least 6 types according to their mass and radius. Two classification algorithms were used: Decision Tree and k-Nearest Neighbor. The latter method, or clustering technique, classifies the multidimensional array of attributes based on resource creation using the mass and the radius of the planet. This in turn creates the special attribute for the classification of Planet Type. In this analysis, both the preliminary validation and the results obtained by the decision tree indicated that there is no correlation between stars and planetary parameters. The classification model returned the results for two dimensions with the following classes: Mercurian, Subterran, Terran, Superterran, Neptunian and Jovian, however did not present results for Mercurian. *As a conclusion, both kNN and Decision Tree algorithms can be used in an ambient with several processing cores, however the kNN algorithm may be clustered. If a high correlation index and low entropy of the data is desired, the kNN may not converge, whereas the Decision Tree presents an instant result based on conditional parameters.*

Keywords: astronomy, classification, data mining, exoplanets.

1. Introduction

To date, about 1,500 exoplanets have been identified, orbiting other stars, and 3,700 planet candidates from the Kepler satellite [1]. Initially,

[1]E-mail Corresponding Author: israel.santos@mackenzie.br

the vast majority of discovered planets were Hot Jupiters, that is, gaseous giants very close to their host star. However, the huge increase in the number of exoplanets in the last years occurred mainly with the Kepler mission, a space probe able to identify even smaller bodies, aimed at finding Earth–like planets within the habitable zone (HZ) of their host star. The habitable zone of a stars is loosely defined as the distance from the star where the temperature on the surface of a solid planet is such that liquid water can exist.

Usually, exoplanets are divided into classes according to their radius and mass. Presently, they are separated into the following types: Asteroid, Mercurian, Subterran, Terran, Superterran, Neptunian and Jovian, as presented in Table 1.

Table 1: Mass classification for solar and extrasolar planets

Planet Type	Mass (Earth Units)	Radius (Earth Units)
Asteroids	0 - 0.00001	0 - 0.03
Mercurians	0.00001 - 0.1	0.03 - 0.7
SubEarth	0.1 - 0.5	0.5 - 1.2
Earth	0.5 - 2	0.8 - 1.9
SuperEarth	2 - 10	1.3 - 3.3
Neptunians	10 - 50	2.1 - 5.7
Jovians	50 - 5000	3.5 - 27

Source: Planetary Habitability Laboratory (PHL)
University of Puerto Rico at Arecibo

Asteroids are small irregular bodies that are not able to maintain a stable environment. Mercurians are able to maintain a significant atmosphere only in cold distant regions, as found on Saturn's largest natural satellite, Titan, the second largest in the entire Solar System and nearly one-and-a-half times the size of the moon. It is the only natural satellite known to have a dense atmosphere, being even denser than Earth's, and the only object to have clear evidence of liquid bodies on its surface, beyond our planet.

Subterran planets are able to maintain a significant atmosphere beyond the outer edges of the habitable zone, being identical to Mars. Terrans can retain a significant atmosphere with liquid water within the habitable zone, identical to Earth. Superterrans are able to sustain dense environments with liquid water within the habitable zone. Netunians, on the other hand, may have dense atmospheres in the hot zone, and finally, Jovians may have super dense environments in the hot zone.

In this context, this study aims to classify all the identified exoplanets,

according to the definition presented by the Laboratory of Habitability Planets [2] of the University of Puerto Rico at Arecibo (Table 1). This may help research on the identification of terrestrial planets that are located within the HZ of their star where life may originate and develop [3].

2. The quest for exoplanets

CoRoT (Convection, Rotation and Planetary Transits) [4] was the first space mission dedicated to exoplanetary research. Launched in December 2006, the mission had a nominal life of 2.5 years, later extended until March 31, 2013, a period that studied the interior of the stars and searched for exoplanets. CoRoT was designed by the CNES (National Center for Space Studies) in partnership with ESA (European Space Agency) and AEB (Brazilian Space Agency).

The Kepler [5] space mission is a NASA-designed (National Aeronautics and Space Administration) observatory to search for extrasolar planets. To this end, the probe observed one hundred thousand stars for a period of four years. The main goal of the Kepler was to search for telluric exoplanets similar to Earth, that is, those between half to twice the size of the Earth. Especially those terrestrial planets within the habitable zone of their host stars, where liquid water may exist in their surface.

The method for planet identification used by both the CoRoT and the Kepler satellites is the detection of small decreases in the light of a star caused by the eclipse or transit of a planet. Planetary transit is the passage of a planet in front of the stellar disk during its orbit, when a dimming of a fraction of a percent in the light of the star can be measured.

The data used in this study was extracted from the *Exoplanet Orbit Database* (EOD) [1]. The content of the database is a compilation of data collected using radial velocity and transit methods for planet detection, including the planets discovered by the CoRoT [4] and Kepler [5] missions. The masses and radius of stars are obtained from precise spectroscopic observations. Moreover, accurate masses of planets are also determined from radial velocity measurements from spectroscopic observation of stars, whereas the radius of the planet can only be measured with transit observations. The amount of dimming in the light of a star during a transit is proportional to the radio of the planet to stellar area (or radius).

Table 2: Quantity of records on the database

Classification	Quantity
Confirmed Planets (EDO)	1491
Kepler Planet Candidates	3704
Total of Planets	5195

Source:Exoplanet Orbit Database [1]

The main classification attributes, that is mass and radius of the planets, are plotted on Figure 1. The data represent planets of different densities, thus explaining the scatter in the figure. Table 2 lists the number of planets in the Exoplanet Orbit Database, whereas

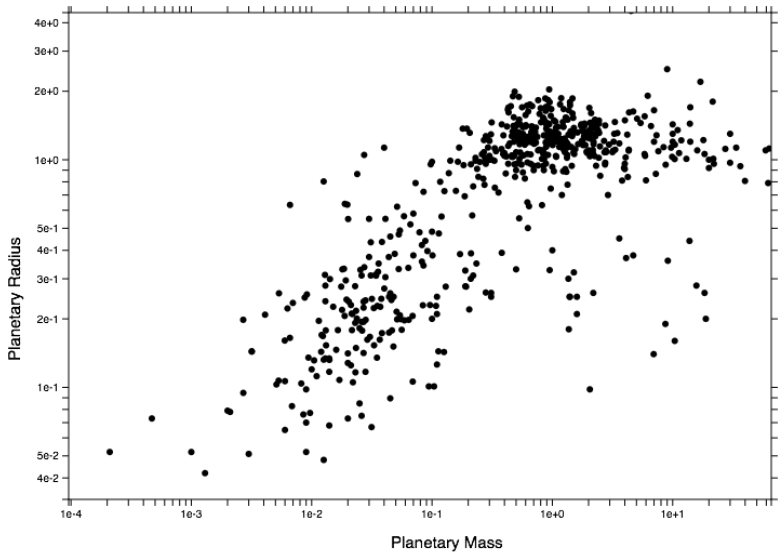

Figure 1: Planet radius (R_Jup) X Planet mass (M_Jup)

Table 3 displays the metadata that characterizes a Star. The text type is defined for the *STAR* attribute and the numeric type is defined for the attributes *MSTAR*, *RSTAR*, *TEFF*, *RHOSTAR*, *LOGG*, *KP*. The planet metadata is shown in Table 4. The numeric type is defined for the attributes *ID*, *PER*, *ECC*, *K*, *T0*, *MSINI*, *A*, *MASS*, *R*. In this work we are manly interested in the *MASS* and *R* parameters.

Table 3: Star Metadata

ID	Name	Unit	Description
STAR	Star Name		Standard name for star
MSTAR	Mass of Star	Solar Mass	Estimated mass of the star, usually based on the association of effective temperature of the star and its luminosity compared to stellar models
RSTAR	Radius of Star	Solar Ray	Estimated Star Ray
TEFF	Teff	Kelvin	Effective Star Temperature
RHOSTA	Density of star	Grass / Centimeter	Density of star as measured from transit photometry and radial velocity information
LOGG	$log10(g)$		$log10$ of gravity on the surface of the star (unit cgs)
KP	KP		magnitude of the Kepler band

Source: *Exoplanet Orbit Database*[1]

Table 4: Planet and Orbit metadata

ID	Name	Unit	Description
PER	Orbital Period	Days	The Orbital Period of the Planet
ECC	Orbital Eccentricity		The eccentricity of the orbit, on the usual scale from 0 to 1, where 0 is circular and 1 is extremely flattened
K	Velocity Semi-amplitude	m/s	The half-amplitude of Doppler variation (half of the variation of the peak-to-peak radial velocity).
T0	Time of Periastron	JD	The time of a planet periastron passage in JD
MSINI	Msin (i)	Jupiter Mass	Minimum Planet Mass. True masses are usually higher by about 15 % due to the geometric inclination effects
A	Semi-Major Axis	AU	Semi-major axis
MASS	Planet Mass	Jupiter Mass	Planet mass
R	Planet Radius	Jupiter Radius	Planet radius

Source: *Exoplanet Orbit Database* [1]

3. Data Mining Concepts

According to Tan, Steinbach and Kumar [6] the input data for the classification are a set of records, each record being known as an instance and

characterized by a double (x, y), where x is the set of attributes and y the special attribute. Classification is the task of learning a target function f that maps each set of attributes x to one of the y labels of pre-determined classes [6]. Also, according these authors a classification technique (or classifier) is a systematic approach to constructing classification models from a set of input data. Among different algorithms, the Hunt algorithm is the basis for many induction algorithms for existing decision trees, including ID3, C4.5 and CART.

According to the Hunt algorithm [6], a decision tree grows recursively by partitioning the training records into successive more pure subsets. Since D_t is the set of training records that are associated with the node t and $y = y1, y2, ..., yc$ are the labels of the classes, we have:

- If all records in D_t belong to the same class y'_t then t is a leaf node labeled y_t.
- If D_t contains records that belong to more than one class, an attribute test condition is selected to partition the records into smaller subsets. A child node is created for each test condition result, and the D_t records are distributed to the children based on the results. The algorithm is then applied recursively to each child node.

There are different approaches to determine the best way to divide the records, the metrics being based on the degree of impurity of the child nodes. As examples are the Entropy (Equation 1), Gini Rate (Equation 2), and Classification Error (Equation 3), among others:

$$\text{Entropy}(s) = -\sum_{i=0}^{c-1} p(i|s) log_2 p(i|s), \tag{1}$$

$$\text{Gini}(s) = 1 - \sum_{i=0}^{c-1} [p(i|s)]^2, \tag{2}$$

$$\text{Classification Error}(s) = \max_i [p(i|s)]. \tag{3}$$

The *k-Nearest Neighbors* (kNN) algorithm [7] stores the training data and when a new object is submitted for classification, the algorithm searches for the closest k records (Euclidean distance) of this new record. Thus, the new record is ranked in the most common class among all the k closest records.

The algorithm applies the concept of centroids. Therefore, for a given a data set, the algorithm randomly selects k records, each representing a cluster. For each remaining record the similarity between the analyzed record

and the center of each grouping is calculated. The object is inserted in the cluster with the shortest Euclidean distance, that is, greater similarity and, with each new element inserted, the centroid is recalculated. Thus, one has:

1. Provide values for centroids
2. Generate a distance matrix between each point and the centroids
3. Place each point in the classes according to its distance from the class centroid
4. Calculate the new centroids for each class
5. Return to step 2 and repeat until convergence

It should be emphasized that different variations implement optimizations for the choice of the k value, measures of dissimilarity and strategies for the calculation of the cluster center. One variation of this is the k-Means [8] algorithm, that uses the mode to calculate the centroids.

4. Planet classification

The clustering technique was used to generate a multidimensional array of *NAME* attributes, the name of the candidate or confirmed planet, *R*, the radius of the candidate or confirmed planet, *MASS*, the mass of candidate or planet, *MSTAR*, the mass of the star, and finally the *RSTAR*, radius of the star. Unfortunately, the Exoplanet Orbit Database [1] has missing data for many records. Thus, for data preprocessing, the data cleaning technique favors the option of excluding such records to avoid inconsistencies in the elaboration of the construction process for later analysis. The technique of resource creation used the mass and the radius of the planet to generate the special attribute for the classification of Planet Type, according to the characteristics presented in Table 1.

The Rapid-Miner® tool 5.3.000 was used to process the data, which was selected from the *Exoplanet Orbit Database* [1], converted to a Microsoft Excel worksheet®, downloaded to the local machine and then submitted to pre-processing. When loading the worksheet to RapidMiner®, heavily typed attributes are used. In this step, the data types of the NAME attributes were defined for *ID*, the *MASS*, *R*, *MSTAR* and *RSTAR* for *NUMERIC*, and finally the *Class* attribute, which was generated by the resource creation technique, has been defined as *LABEL*.

The parametrization criteria of the algorithms were empirically tested for the software, algorithm and specificities that make up the Data Mining

technique. Finally, the preliminary validation of the data was performed based on the metric presented in Table 1.

We applied the two algorithms described in Section 3 to classify the planets into Planet Types. First, the Decision Tree algorithm was used. Figure 2 presents the result of sorting by the decision tree according to *MASS*, using the Gini index. Sorting by the decision tree using the information gain yields the same results. Therefore, no change in the structure of the trees with the index exchange occurs, basically because the initial parameters were not altered. The decision tree algorithm applied to the planet radius attribute R yields the same results, and thus are not shown here.

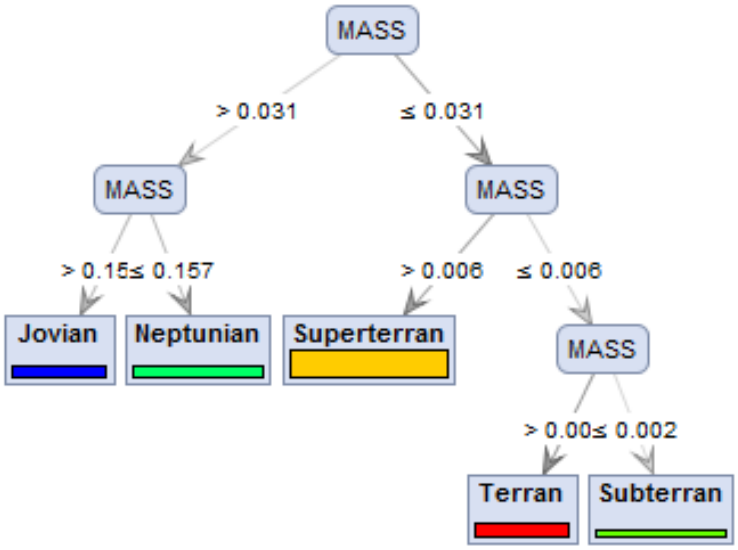

Figure 2: Decision tree for MASS attribute.

Next, the second algorithm, kNN, was applied to the planet database using both the planet mass and radius attributes. Table 5 presents the first clustering attempt using the *k-Means* algorithm without the special attribute, where it can be seen that the k variation generated groups in border areas.

Table 5: k-Means with 6 Clusters

Cluster	items
0	544
1	15
2	767
3	207
4	13
5	1035
6	181
Total number of items	2762

Considering the initial planet types given in Table 1, the algorithm matches the clustering data of Table 5 into the predicted planet classes, that yields the accuracy listed on Table 6. The classification model returned the result for two dimensions with the following classes: Mercurian, Subterran, Terran, Superterran, Neptunian and Jovian, as summarised in Table 6. As can be seen from the values presented on the Table, the kNN algorithm show accuracies above 80% in most of the cases.

Table 6: K-Means Performance for 6 Clusters

Prediction Class	Accuracy						
	Merc.	SubT.	Terran	SuperT.	Nept.	Jovian	Precision%
Mer.	0	0	0	0	0	0	0.00%
SubTer.	2	44	15	0	0	0	72.13%
Ter	0	31	384	47	0	0	83.12%
SupTer.	0	1	60	1434	44	0	93.18%
Nep.	0	0	0	23	323	7	91.50%
Jov.	0	0	0	0	4	343	98.85%
%	0.00	57.89	83.66	95.35	87.06	98.00	

For internal validation of the k-NN algorithm, the number of initial clusters was increased to 12. Doubling the number of clusters to 12 does not alter significantly the results, thus validating the method.

5. Conclusions

Two classification algorithms were used to classify almost three thousand planets in the Exoplanet Orbit Database (EOD), according the their mass and radius (Table 1). The initial classification performed in this work, will help identify terrestrial planets within the HZ of their host star, since life as we know it needs liquid water on the surface of a terrestrial planet to develop.

Both algorithms, Decision Tree and kNN, have contributed in some way to the validation or clustering of data. However, the performance between the algorithms tends to become computationally costly depending on the initial configurations. The decision tree algorithm required a preprocessing of the data to ensure that there was no absence of values, in the definition of the choice of a special attribute for the classification of the samples, whereas the kNN proved to be an algorithm that has supervised learning based on the instances present in the data mass. Thus, it is not possible to choose a special class attribute in this algorithm because the comparison is performed in relation to all other known copies in the knowledge base, and this comparison can consume a lot of computing time. In particular, the Rapid-Miner® tool used in the present study allowed a more accurate analysis of the data through different algorithms.

The current paradigm of science consists in allying the use of computational techniques with the other areas of knowledge, such as Astronomy. More specifically, the results obtained from the planet classification scheme presented here constrained the number of terrestrial planets candidates, thus making it possible to focus on the set of life bearing planets.

Acknowledgments. Israel dos Santos acknowledges financial support from MackPesquisa, A.V. thanks FAPESP for partial financial support, and A.H.F.G. acknowledges a fellowship from CAPES.

References

[1] Exoplanet Data Explorer. Available: http://www.exoplanets.org

[2] Wright, J. T., Fakhouri, O., Marcy, G. W., Han, E., Feng, Y., Johnson, J. A., ... & Piskunov, N. (2011). The exoplanet orbit database. Publications of the Astronomical Society of the Pacific, 123(902), 412-422. DOI: 10.1086/659427

[3] Holman, M. J., & Wiegert, P. A. (1999). Long-term stability of planets in binary systems. The Astronomical Journal, 117(1), 621. DOI: 10.1086/300695

[4] Astronomy mission. From stars to habitable planets. Available: http://smsc.cnes.fr/COROT/

[5] Kepler. A search for Habitable Planets. Available: http://kepler.nasa.gov/

[6] Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. In Library of congress (Vol. 74).

[7] McNames, J. (2001). A fast nearest-neighbor algorithm based on a principal axis search tree. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(9), 964-976.
DOI:10.1109/34.955110

[8] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.
DOI: 10.2307/2346830