Boletim da Sociedade Astronômica Brasileira, **31**, no. 1, 21-22

Classification of exoplanets with data mining techniques

Gabriel M. Manfredi[1], Henrique S. Areias[1], Natale L. Pupo[1], I.F. Santos[1], & Adriana Valio[2]

[1] Mackenzie Presbyterian University, Computing and Informatics Faculty
e-mail: gabrielmmanfredi@gmail.com, henriquesareias@hotmail.com, leobrasil@gmail.com, learsi.isr@gmail.com

[2] Mackenzie Presbyterian University, Center for Radio Astronomy and Astrophysics at Mackenzie; e-mail: adrivalio@gmail.com

Abstract. This work presents the use of the K-means algorithm for the classification of exoplanets discovered by the Kepler mission through planetary transit. The goal is to identify planets that are in the habitable zone of the host star and, therefore, more likely to have water in the liquid form on its surface. These planets would be the best candidates for the existence of life in the form that we know it.

Resumo. Este trabalho apresenta o uso do algoritmo K-means para a classificação de exoplanetas descobertos pelo satélite Kepler por meio de trânsito planetário. O objetivo é identificar planetas que se encontram na zona habitável da estrela hospedeira e, portanto, com maior probabilidade de se encontrar água na forma líquida. Consequentemente, estes planetas seriam os melhores candidatos para a possibilidade da existência de vida na forma que a conhecemos.

Keywords. Stars: statistics

1. Introduction

Astronomy research uses computational resources to support scientific findings, due to the large volume of data resulting from processing of any equipment used to collect this type of data. In this context, according to Stefanowski (2009), data mining is a computational resource that aggregates traditional methods of data analysis with sophisticated algorithms in the processing of these large volumes of captured data. This work aims to automatically classify exoplanets using the data from the Exoplanet Orbit Database (EOD) (Wright et. al. 2011) and applying data mining techniques.

2. Data Mining

Data Mining refers to the discovery of new information due to patterns in large amounts of data (Fayyad et. al. 1996). Thus, Knowledge Discovery in Databases is defined as the discovery of knowledge of the data, while Data Mining refers to the application of algorithms for the extraction of Fayyad et. al. (1996) models. Sample data can be observed in the Table 1 and Figure 1 present the steps used in this process.

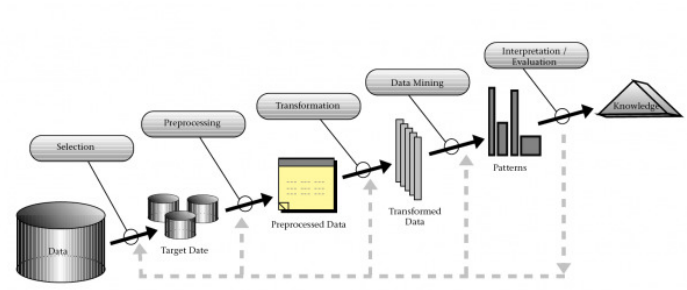

Figure 1. Process Steps Knowledge Discovery in Databases (KDD).

Table 1. Extract of available data from the Kepler Mission. Source:http://www.exoplanets.org/table

Name	Mass of Star	Radius of Star	Planetary Radius	Planet Mass	Distance Star
Kepler 107 d	0.510	1.411	0.0955	0.00371	129.0
Kepler 1049 b	0.950	0.490	0.085	0.00245	840
Kepler 813 b	0.960	0.93	0.191	0.0160	1100

Table 2. Classification of planets based on radius and mass. Source: Planetary Habitability Laboratory (PHL) Puerto Rico University, Arecibo

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	0.2 - 10	1.3 - 3.3
Neptunians	10 - 50	2.1 - 5.7
Jovians	50 - 5000	3.5 - 27

3. Exoplanets

The Kepler mission monitored about 160,000 stars in the direction of Cygnus, searching for the small signatures caused by the transit of planets in orbit of their host stars (Chaplin et. al. 2011). The classification performed here is based on the attributes of mass and radius of the planets, in units of Earth mass and radius. Thus, the planet type, based on its mass and radius, is taken from the exoplanets groups presented in Table 2.

4. Data Mining Techniques: K-means Algorithm

K-means is an algorithm that applies the centroid concept and uses an efficient method based on the concept of Euclidean distance. The goal is to find the similarity between the data and

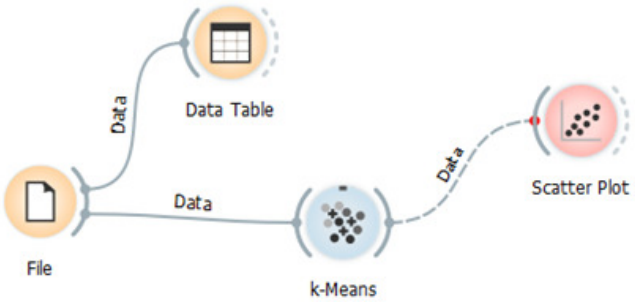

Figure 2. K-means execution conception.

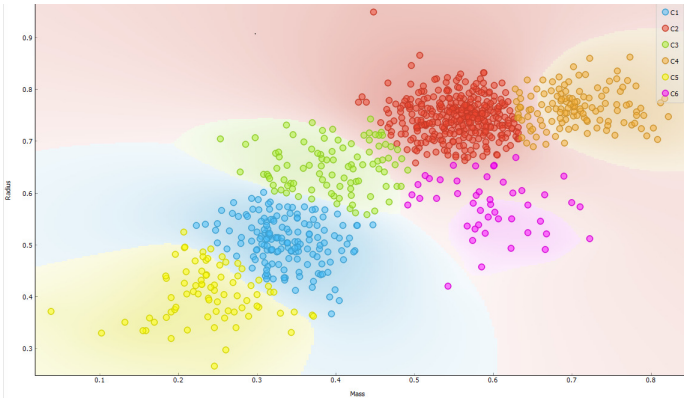

Figure 3. Result of the K-means application.

group them according to the defined value of the k argument, which is defined by the iterative execution of the algorithm. The disadvantage of using the algorithm is the choice of the initial value k, since a too small number of sets can generate the merger of two natural clusters, while the choice of a large number can generate the separation of a natural set in two.

5. Classification of Exoplanets using the K-means algorithm

The data obtained from the Kepler mission required preprocessing and cleaning, since it is essential to eliminate redundant and inconsistent data, as shown in Figure 3.

The data were converted to the appropriate use of the algorithm, for example, integers and decimals. Finally, the application of K-means was done by submitting the data set as parameter to the algorithm. The configuration of the algorithm parameters and the plot after classification were obtained considering the parameter $k = 6$, based on the classification of Arecibo (Table 2).

The groups were classified as exoplanets based on their radius and mass and the result is shown in Yellow (Mercurians), Blue (Sub Earth), Green (Earth), Purple (Super Earth), Red (Netunian) and Orange (Jupiter), accordant showed Table 2.

6. Conclusions

The classification of the exoplanets was performed using the k-means algorithm and the result presented similarity with the data found in the literature. In this context, the next step will be to use the Kepler light curves in order to search only the exoplanets classified as Earth-like for model extraction.

Acknowledgements. We thank MackPesquisa, the research funding programme of the Mackenzie Presbyterian Institute and FAPESP (São Paulo Research Foundation)

References

Stefanowski, J. (2009). Data Mining-Clustering, University of Technology, Poland

Wright, J. T., Fakhouri, O., Marcy, G. W., et al. 2011, PASP, 902, 412

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. 1996, Knowledge Discovery and Data Mining: Towards a Unifying Framework, in KDD (Vol. 96, pp. 82-88)

Bernstein, P. A., Hadzilacos, V., & Goodman, N. 1987, Concurrency control and recovery in database systems

Han, J., Pei, J., & Kamber, M. 2011, Data mining: concepts and techniques. Elsevier

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992), Knowledge discovery in databases: An overview, AI magazine, 13, 57.

Chaplin, W. J., Kjeldsen, H., Christensen-Dalsgaard, J., Basu, S., et al. 2011, Science, 6026, 213