VOL.5 NO.2 September 2016 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Comparison of expectation-maximization clustering and logistic regression on categorization of planets with known properties

Ajla Suljević Pašić, Sadina Gagula-Palalić

International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
HrasnickaCesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina

ajla.suljevic_pasic@yahoo.com; sadina@ius.edu.ba

Article Info

Article history:
Article received on 7 Jun. 2016
Received in revised form 2 Sep.2016

Keywords:
Exoplanets, Categorizing, Comparison, Expectation-maximization clustering, Logistic Regression, Machine Learning.

Abstract

Analysis of the exoplanet data is the top priority of astrophysicists today. With the increasing incoming information there is a need for an efficient and reliable algorithm. The data is taken from exoplanet data explorer which was cross checked and filtered with NASA's known categorization. These were then sorted into 5 categories: Dwarfs, Terrestrial, Icy, Jovian and Giant planets. This paper compares expectation-maximization clustering algorithm as an unsupervised and logistic regression as a supervised machine learning methodologies. Comparatively, logistic regression outperformed EM, indicating it cannot be used to sort through the incoming data. Further analysis is necessary.

1. INTRODUCTION

Exoplanets are those celestial bodies that circle stars other than our Sun. They do not shine, are small comparing to their starry neighborhood, and far away from our observational point. That is why, even though theoretically they were a certainty, it was only recently that technology was developed in order to detect them. [1][2] At this moment there is 3280 confirmed, and 2416 possible planets detected. [3]

The new data is collected and regularly published on a website containing the database with as many features as the telescopes and observatories can collect, both, of the planets and their host stars. The database is easy to explore, it is possible to run various queries online, as well as select what features to observe. Each feature comes with an explanation and units used. [4]

With the emergence of fast and efficient exoplanet data collection technologies, the number of observations has been accelerating. In 2005 there were 152 detections over 10 years period. These data were easy to manually examine and confirm. [5] At this moment about 1000 new planetary candidates are discovered each month. [1][4] This is why there is an urgent need to find a best algorithm to sort through it efficiently and accurately. [6]

The decision boundaries and relations between parameters are not always clear, so manual analysis takes too much time, even with teams of analysts worldwide. [7] Further problem is that there is not a set of standardized values that would help scientist categorize the data more efficiently. [4] The limits and categories, as well as the defining features, vary from research to research. This paper categorizes planets as proposed by D. G. Russell in *Solar and Extrasolar Planet Taxonomy* into: dwarfs, terrestrial, mixture or ice, jovian, and giants. [8]

In order to preprocess the data and apply algorithms, data mining software in Java, WEKA 3.8, was used. This is

University of Waikato open source product issued under GNU General Public License. The explorer allows for a fast and easy, graphical and numerical representation of the data. It provides an overview of the algorithms that can be applied to the given data, as well as outputs, testing information and statistical breakdown of the results. [9]

The clear categorization is required to assess performance between expectation-maximization clustering as an unsupervised and logistic regression as a supervised, probabilistic machine learning methodologies discussed hereon.

2. DATA

The data was taken from exoplanets.org on 10th May 2016. The table, at that moment had about 50 features out of which 12 were generally known to the wider public and therefore selected for this analysis. These were then decreased to 7 in preprocessing. The chosen features were: mass, radius, density, and gravity of the planet, separation from the host star, star's mass, and its radius. [10]

Number of samples available was 2139. Out of that 581 samples were with most data available, namely with ≤ 3 zeroes per row. This is including the planets in our solar system. [10]

The next step was to find output information. The individual planet classification table reliably sourced is not available. Therefore, it was necessary to sort through known and available data to construct this feature.

2.1 Dwarfs

Dwarfs are small celestial objects that have enough mass to have circular shape and orbit the sun. They are defined as bodies whose mass is below 0.0002 in respect to the mass of Jupiter, as defined by Russell's taxonomy. [8]

Dwarf planets are common but difficult to discover due to their small size. Only 10 of them were within the given data, out of which 5 are in our own solar system. They are categorized as 0.

2.2 Terrestrial

Regardless of the actual chemical composition of the ground and atmosphere, terrestrial planets are all whose ground is rocky. In our solar system first four planets: Mercury, Venus, Earth and Mars are of this category.

Furthermore, the terrestrial planets are studied in details as they are considered potentially habitable, based on our planet. This list excludes some of the rocky planets; nevertheless, it can serve the purpose for this analysis. The list provides 42 objects. [11] However, it doesn't include all objects with clear characteristics of rocky worlds, so the extended list includes potential candidates, adding up to 79 samples. The limiting parameter is the mass of the planet which should be between .0002 and 0.02 of Jupiter's mass [8]. They are categorized as 1.

2.3 Icy or uncategorized

The third category or #2 in the table are dense gaseous, liquid, ice, and/or large rock planets. There were 42 such samples. Uranus and Neptune in our solar system fall within this category. This is the most complex group. Our solar system doesn't have many examples or possibilities of planets within this category. However, it has been discovered that some of the planets within the scope are not even theoretically predicted. [2] The taxonomy defines them as the class between jovian and terrestrial bodies, mass between 0.02 and 0.08 of the Jupiter's. [8] This in no way defines their structure, composition, position or characteristics, unlike other categories. This group can, therefore, be considered as uncategorized celestial bodies.

2.4 Jovian

Jovian or gaseous planets, class 3 in this categorization, are in well-defined parameters in "*Solar and Extrasolar Planet Taxonomy"* so 230 of them were categorized. [8] Jupiter and Saturn are the examples of this category within our own solar system. In fact, it is the Jupiter that gave this group its name. This group is characterized by its own thermal radiation, which can be further used to confirm detections. [1]

2.3 Giants

Giant is a descriptive term. It describes all the planets whose mass surpasses 6 relative to the Jupiter. However, this group takes only giant gaseous planets in consideration. There are Giant rocky, icy or even liquid worlds. They are not accounted here as their characteristics largely differ from one another. The gaseous planets of this mass usually have some percentage of H/He conversion. However if that conversion is significant those bodies are categorized as brown dwarfs, or the smallest class of stars. Therefore this category takes into consideration only those bodies whose mass doesn't surpass 42 masses of Jupiter. [8] At the time of this analysis, in the fourth category 11 of them were classified as such, notably with majority data known. [10]

In conclusion, the final version of data table had 372 samples with known outputs. The mass-radius distribution of the data is shown below, and it has Gaussian-like shape.

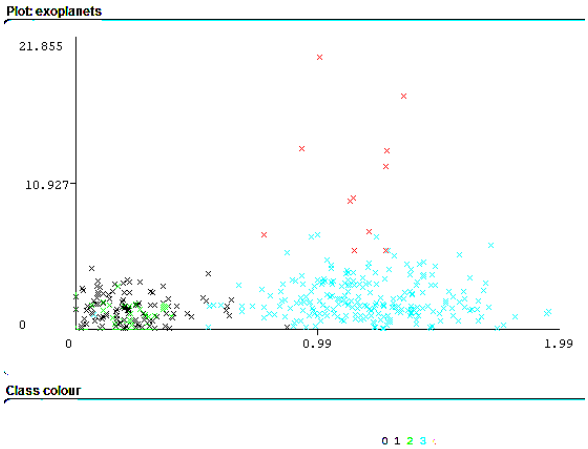

Figure 1: Radii of planets are on horizontal axes and masses of planets are on vertical axes

3. CLUSTERRING USING EXPECTATION-MAXIMIZATION ALGORITHM

The main focus of this ongoing research is to find best unsupervised learning algorithm in order to sort through new incoming data quickly and efficiently. EM algorithm or expectation-maximization, as an iterative method, is good for latent or hidden variable problems – unknown connections as well as missing data problems. It is usually used for exponential families. However, it has proven to be the best clustering model, for organizing planets scattered with Gaussian distribution. E-M maximizes probability of known data by iteratively improving coefficients of the expected values –both known and unknown. [12]

Namely, an estimation problem with $\{x^{(1)}, \ldots, x^{(m)}\}$ training set comprised of m examples is used to fit the parameters of a model p(x, z) to the data. The z here is all the unknown dimensions of the data. The likelihood of the output is:

$$l(\theta) = \sum_{i=1}^{m} \log p(x;\theta)$$

$$l(\theta) = \sum_{i=1}^{m} \log \sum_{z} (x, z;\theta)$$

The next step is maximizing $l(\theta)$. This is difficult to do, especially in many feature sets. The applied strategy is to iteratively construct a lower-bound on l (E-step), followed by optimization of that lower-bound (M-step). This process is repeated until convergence. [12]

Therefore, the EM algorithm is as follows:

Repeat until convergence: {

E step: $Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\ \theta)$

M step: $\theta := arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}$

}, where Q is some distribution over z's.[12]

The EM algorithm was applied using WEKA software for data mining tasks.

4. CLASSIFICATION USING LOGISTIC REGRESSION

The supervised machine learning algorithm chosen as a reference is multiclass logistic regression. This is, also, a probabilistic approach for the data analysis. It is considered to be a faster way of data mining to the Gaussian models. [13]

Logistic regression of a two class case is defined as the posteriori probability of a class *C* as a sigmoid acting on a linear function of the feature column ϕ:

$$p(C|\phi) = y(\phi) = \sigma(w^T\phi)$$

This means that the probability of a sample to be characterized within C category is a sigmoid of transposed weights of the given feature matrix. [13]

The multiclass logistic regression model, however, is more complex:

$$p(C_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

Where activations are:

$$a_k = w_k^T\phi$$

The maximum likelihood to determine separable characteristics of the classes and their densities is implicitly determined here. [13] Since this is not an iterative method, at least without back propagation to decrease the error, this is sufficient to sort the samples into distinct categories. The logistic regression was also applied using WEKA software for data mining tasks.

5. COMPARISON AND CONCLUSION

The table below shows the output feature, numerically and percentage-wise. Comparatively, it outlines the categorization of 15% of the same data tested with both, clustering and classification.

Type	Known		Expectation-Maximization		Logistic Regression	
Dwarf	1	2%	1	2%	1	2%
Terrestrial	13	23%	8	14%	12	21%
Icy	6	11%	11	20%	7	13%
Jovian	33	59%	30	54%	33	59%
Giants	3	5%	6	11%	3	5%
Total	56	100%	56	101%	56	100%

Table 1: Tested planets' categories, comparatively in numbers and percentages for known planets, and Expectation-Maximization and Logistic Regression results

As it is evident, there is a large difference between clustering that was unsupervised and the regression which

used outputs to learn how to separate terrestrial and icy planets. Furthermore it appears that some of the Jovian planets were categorized as giants using EM algorithm.

Consequently, the overall overview comparison is as follows:

	Expectation-Maximization	Logistic Regression
Preprocessing	Yes	Yes
Execution time	2.86 s	0.52 s
Iterations	4	NA
Accuracy	71%	93%
Classification error	29%	7%

Table 2: Comparing the results of Expectation-Maximization and Logistic Regression results

Considering that the taxonomy is not a natural and self-evident occurrence, but manmade limitations to assist in distinguishing between different objects, it is hardly a surprise that the algorithm which performed better is the one that had the information of the named distinctions. At nearly 93% accuracy and a half a second to perform, logistic regression is the best algorithm for these data.

Expectation-Maximization, even though theoretically is a good fit, has not shown to be adequate mistakenly categorizing a third of the data into wrong clusters. Subsequently, either a variation of the algorithm or an entirely new methodology should be used to analyze the exoplanet data.

However, further studies must be conducted before any strong conclusion can be made. More features or even more samples should be taken into advisement for conclusive assessment of EM algorithm.

REFERENCES

1. Y. Sun, S. Ferraz-Mello, J. Zhou,*Exoplanets: Detection, Formation and Dynamics*, International Astronomical Union Symposium No. 249, 2007, DOI: 10.1017/S1743921308016608
2. C. D. Dressing, *New Frontiers in Exoplanet Detection: High Contrast Imaging with Subaru*, Princeton University, 2010
3. P. Brennan, *NASA's Exoplanet Exploration Program,* California Institute of Technology, 12 June 16, https://exoplanets.jpl.nasa.gov/
4. J. Schneider, C. Dedieu, P. Le Sidaner, R. Savalle, and I. Zolotukhin, *Defining and cataloging exoplanets: the exoplanet.eu database,* Astronomy & Astrophysics no. 532, A79, 2011, DOI: 10.1051/0004-6361/201116713
5. G. Marcy, et al, *Observed Properties of Exoplanets: Masses, Orbits, and Metallicities,* Progress of Theoretical Physics Supplement No. 158, 2005
6. R. L. Akeson, X. Chen, D. Ciardi, M. Crane, et al *The NASA Exoplanet Archive: Data and Tools for Exoplanet Research* Publications of the Astronomical Society of the Pacific 125.930 (2013): 989-99.
7. D. A. Fischer, A. W. Howard, et al, *Exoplanet Detection Techniques*, Protostars & Planets VI, Convention Center Heidelberg, Germany, 2013
8. D. G. Russell., *Solar and Extrasolar Planet Taxonomy,* Owego Free Academy, 2013.
9. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *The WEKA Data Mining Software: An Update*, SIGKDD Explorations, Volume 11, Issue 1, 2009
10. J. Schneider, *NASA Exoplanet Archive,* University of Massachusetts and IPAC/Caltech collaboration, 10 May '16, https://exoplanets.org/
11. A. Mendez, *The Planetary Habitability Laboratory,* University of Puerto Rico at Arecibo, website, 10 May '16, http://phl.upr.edu/projects/habitable-exoplanets-catalog
12. A. Ng, *The EM Algorithm,* Stanford University, 2016
13. C. M. Bishop, *Pattern Recognition and Machine Learning* , Springer, 2006, ISBN-10: 0-387-31073-8