VOL.6 NO.2 September 2017 - 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

Categorizing Stars with Known Properties Using the Expectation-Maximization Clustering Algorithm

Ajla Suljevic Pasic, Assist. Prof. Dr. Emine Yaman
International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
Hrasnicka Cesta 15, Ilidza 71210 Sarajevo, Bosnia and Herzegovina
ajla.suljevic_pasic@yahoo.com; eyaman@ius.edu.ba

Article Info

Article history:

Article received on 13 June 2017
Received in revised form 1 August 2017

Keywords:

Stars, Categorizing, Hertzsprung-Russell diagram, Comparison, Expectation-maximization clustering, Machine Learning.

Abstract

Hertzsprung and Russell, created a diagram of then known stars with respect to absolute magnitudes or luminosities versus their stellar classifications or effective temperatures. This gave a clear clusters of star types, namely main sequence stars, from birth to maturity, followed by giants, supergiants and white dwarfs. With the rise of technology number of stars with known properties had been growing exponentially and manual categorization is futile. Using the same parameters of HR diagram, this paper analyzes the efficiency of unsupervised ML algorithm expectation-maximization clustering on a database containing 120 000 stars.

1. INTRODUCTION

The first systematic classification of stars date back to the beginning of 20th century. Hertzsprung and Russell created a diagram of then known stars with respect to absolute magnitudes or luminosities versus their stellar classifications or effective temperatures. It also shows the evolution of star's life, from birth to death. [1] The diagram, as it can be seen in Figure 1, showed that all stars can be categorized into 4 clusters based on the type of star compared to our Sun:

1. Main sequence contains stars from early stages upon formation to near end-of-life-maturity (Sun is a part of the main sequence). Approximately 90% of stars fall within this group. The key characteristic of these stars is that they are all fusing hydrogen in their cores.
2. Giants are the mature stars. They not only fuse helium but also burn hydrogen in a shell around the core. Once they die they become the white dwarfs.
3. Supergiants are stars with the most matter that didn't turn into white dwarfs but continued growing after giant stage. These stars explode at the end of their lives into supernovas and turn into black holes.
4. White dwarfs are remnants or corpses of the dead stars. Their shine comes from the remainder energy of the cores of giants. [2]

Black holes are the ultimate remnants of stars. They are not on the graph since they do not emit light. In fact, they absorb it. [3] This article, therefore, does not address them.

Colors of stars depend on their chemical composition, which depends on the stars formation and stage of life. [2]

The paper shows how unsupervised machine learning algorithm, namely, expectation-maximization clustering categorizes stars comparing to the Hertzsprung-Russell diagram.

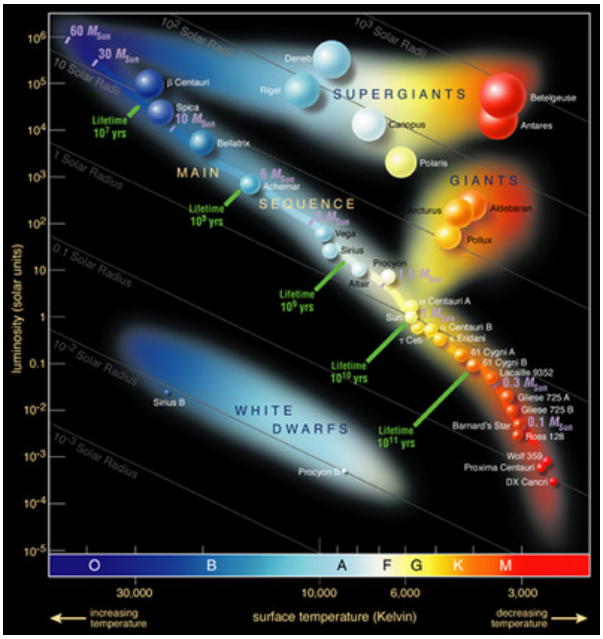

Figure 1: colorized HR diagram

2. DATA

The data was taken on the 21st of May 2017 from The Astronomy Nexus, privately owned and opened for public use collection of different astronomical databases. The database contained all stars in Hipparcos, Yale Bright Star, and Gliese catalogs. [4]

Hipparcos catalog was a result of European Space Agency mission in mid 1990s that recorded data for 4 years. It summarizes main astrometric and photometric properties of 118 218 stars. [5]

Yale Bright Star catalog (BSC) is public database stars visible to the naked-eye containing basic astronomical and astrophysical data. In addition to the collected numeric data, catalog contains information on individual entries, such as star names, colors, companions, constellations, types, and many other known stats. The catalogue has 9096 stars. [6]

Gliese catalog of Nearby Stars focuses on stellar objects within 25 parsecs of the sun. It contains details on 3803 stars with precise quantifiable data and additional detailed information, such as color, size, and type. [7]

The table, at that moment had about 36 features out of which only radius, distance, stellar classification,

magnitude, and luminosity are selected in preprocessing (Table 1). [4]

Number of samples available was 119 615, including our Sun. All fields were filled with relevant information, except the stellar classification. There are 3 051 stars that do not have this information. [4]

Expectation maximization algorithm works even when there is data missing, [8] therefore these stars were not remove from the database.

Table 1: Database example

ra	dec	dist	mag	abs m	spect	lum
0	0	0	-26.7	4.850	G2V	1.00
0.0001	1.1	2.1978	9.10	2.390	F5	9.6383
0.0003	-19.5	479.62	9.27	5.866	K3V	0.3923
0.0003	38.9	4.4248	6.61	-1.619	B9	3.8690
0.0006	-51.9	1.3422	8.06	2.421	F0V	9.3669

3. CLUSTERRING USING EXPECTATION-MAXIMIZATION ALGORITHM

The main aim of this research is to find out if unsupervised learning algorithm can sort through new incoming stellar data reliably, quickly, and efficiently. Expectation-maximization (EM) algorithm is an iterative method that clusters the entries of the fed database based on probability of entry belonging to a cluster. It is particularly efficient with latent or hidden variable problems. Algorithm is effective in dealing with unknown or imperceptible connections as well as missing data problems. [8]

It is typically used for exponential families. However, it has proven to be the best clustering model, for organizing large amounts of data with no precise parameter limitations that separates different categories. EM maximizes probability of known data by iteratively improving coefficients of the expected known and unknown values. [8]

Expectation step (E) takes a training set of m examples $\{x^{(1)}, \dots, x^{(m)}\}$ that is estimated and fit into the parameters of a model p(x, z). Variable z represents all the unknown dimensions of the data. In our case, stellar classification of the 3k examples. Statistically, 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 (M) the likelihood $l(\theta)$. Since the dataset of this research doesn't have many features, this step doesn't take a lot of time. However, for larger sets maximizing the likelihood takes away the most resources. This step cannot be separated from E step since it is iteratively constructing a lower-bound on l (E-step), followed by optimization of that lower-bound (M-step). The method is reiterated until convergence. [8]

Thus, the EM algorithm is:

Repeat until convergence: {

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

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

where Q is some distribution over z's. [8]

4. IMPLEMENTATION

The EM algorithm was realized using the Weka software. Weka 3.8 is a data mining software in Java created at the University of Waikato. It is open source product issued under GNU, General Public License. [9]

The program described expectation-maximization as a simple probability distribution for instances belonging to each cluster. Clusters are determined either by cross-validation or specified apriori. [9]

The issue of missing numeric values was solved by global replacement, assigning them mean and mode values from the training data. Non-numeric data are assigned null. [9]

There are several parameters that can be determined prior to the execution. They are preset for maximum efficiency in most executions. Those values are:

1. Debug (default F) – additional cluster information
2. Display model in old format (default F) – old format is better for higher number of clusters
3. Do not check capabilities (default F) – reduces run time, but can interfere with cluster determination
4. Maximum iterations (default 100)
5. Maximum number of clusters (default -1) – best left to the algorithm to determine

6. Minimum log likelihood improvement CV (default 10^{-6}) – cross-validation probability improvement needed to increase number of clusters
7. Minimum log likelihood improvement iterating (default 10^{-6}) – minimum improvement essential to complete another iteration of EM steps
8. Minimum standard deviation (default 10^{-6}) – minimum allowable
9. Number of clusters (default -1) – must be equal or less than maximum
10. Number of execution slots (default 1) – set to the number of CPU cores available
11. Number of folds (default 10) – used for cross-validating to find best number of clusters
12. Number of K means runs (default 10)
13. Seed (default 100) – randomly used seed number [9]

5. RESULTS AND DISCUSSION

The table below shows the resulting clusters, numerically and percentage-wise, as well as comparative theoretical HR stats.

Table 2: Tested stars' categories, comparatively in numbers and percentages for statistical theoretical data, and Expectation-Maximization results

Type	HR [10]		Expectation-Maximization	
	#	%	#	%
Main Sequence	107652	90	60597	51
Giants	478	0.4	3949	3
Super Giants	35	0.0003	10215	9
White Dwarfs	10765	9	44853	37
Total	118930	99.4	119614	100

As it is evident, there is a large difference between clustering that was unsupervised and theoretical model used in practice. Stars in their infancy have also been categorized as dwarfs. Stars that are a part of the main sequence nearing the end of life appear to be Giants and Supergiants. This might appear to be correct if only magnitude and radius of stars was observed. However, other parameters are used to select between those features to help improve accuracy and near the effectiveness of HR model.

Table 3: Statistics of the EM performance

	EM algorithm
Number of samples	119 614
Time taken	794.75 s
Log likelihood	-37.5952
Iterations	1

The algorithm is fast in respect to the number of items processed and it took only 1 iteration to reach the efficiency of 10^{-6} for the cluster sets. However, as it can be seen on figure 3, it has failed to reach HR standards.

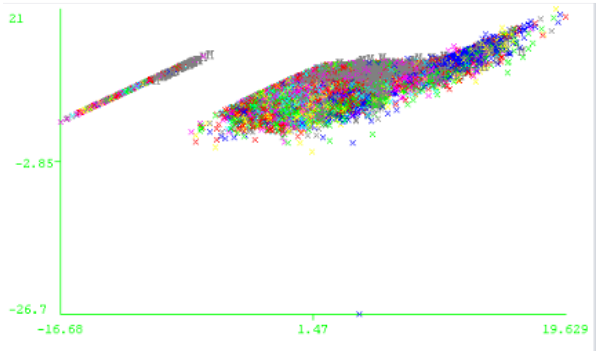

Figure 2: Lum/mag vs. stellar type graph [9]

Since the Hertzsprung and Russell created a diagram of then known stars and the clear pattern of categories showing stars' size, age, color, and temperature has emerged, the nomenclature and categorization of stars in official astronomical circles has heavily relied upon it. Throughout the century many attempted variations on the topic, but the majority has fallen back to their work as astronomical observational standard. With advancement of the technology and exponential increase in statistics of the stellar objects, nearby and in the deep space, the HR diagram received only higher accuracy. [1]

In recent years, there were several attempts to use machine learning algorithms, both supervised and unsupervised, to achieve same or better results. [11, 12, 13] In *Data Mining and Machine Learning in Astronomy,* authors conclude that ML is useful only for solving specific problems, rather than being applied to broad sets of objects. [11]Same authors demonstrated the claim in the article *Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees* – where they analyzed 143 unique photometric objects to have 15% of them misclassified using the decision trees. [12]

S. Jing, W. DeSheng, and L. GuangRui in the article *An Efficient Guide Stars Classification Algorithm via Support Vector Machines*, demonstrated that even supervised algorithm has fallen behind the existing model despite showing a supreme performance comparing to then existing ML algorithms. [13]

Consequently, the fact that EM has failed to reach HR statistical accuracy and efficiency is not unforeseen.

6. CONCLUSION

Expectation-Maximization is theoretically a good fit for this problem: it can efficiently process large quantity of data and maximize statistical probability of an entry belonging to a cluster, even when not all data is available. It has recognized that there are 4 clusters that stellar objects belong to. However, it has not shown to be adequate since it mistakenly categorized majority of the data into wrong clusters.

The aim of the study was to show that unsupervised algorithms can precisely fit data into select groups. Despite recognizing the 4 groups efficiently, EM has failed to precisely sort the data.

1. FUTURE STEPS

Nevertheless, further studies must be conducted before any strong conclusion can be made. The fact that algorithm determined that 4 clusters are the best solution to the stellar classification is promising.

Before anything else, running the same algorithm on larger set with more features might give better results within the current frames. Furthermore, adjusting the algorithm parameters to non-default values, that is, tweaking the algorithm on that set, whilst providing more computing resources, appears to be the right way forward.

Another way to go is to consider chemical compositions of stars, since they indicate type, color, radius, mass, and age among many characteristics. This can be achieved using not only above catalogues but more concurrent stellar datasets.

These cases, however, are not considered for the purposes of this paper. The article focuses on analyzing the effectiveness of unsupervised machine learning algorithm on a database containing information that was available over 100 years ago. The results, even though they are not equal to manually obtained clusters, do show some of the same characteristics and therefore encourage further investigations into the topic.

REFERENCES

1. L. A Hillenbrand, A. Bauermeister, R. J. White,*An Assessment of HR Diagram Constraints on Ages and Age Spreads in Star-Forming Regions and Young Clusters*, Cool Stars, Stellar Systems, and the Sun XIV, ASP Conference Series, Vol. XX, 2007
2. M. Luciuk, *The HR Diagram -the Most Famous Diagram in Astronomy,* Taken on: 4th June 2017 http://www.asterism.org/tutorials/tut39%20HR%20Diagram.pdf
3. *What Is a Black Hole?,*NASA Knows, Taken on: 4th June 2017

   https://www.nasa.gov/audience/forstudents/5-8/features/nasa-knows/what-is-a-black-hole-58.html
4. *D. Nash, The HYG Database*, v3, update 2015 http://www.astronexus.com/hyg
5. M.A.C. Perryman, L. Lindegren, J. Kovalevsky, E. Hog, et.al., *The Hipparcos Catalogue*, ESA, A&A...323L..49P, 1997, Taken on 6th May 2017 http://adsabs.harvard.edu/cgi-bin/bib_query?1997A&A...323L..49P
6. D. Hoffleit, W. H. Warren Jr., Yale *Bright Star catalog*,5th Revised Ed, Astronomical Data Center, NSSDC/ADC, 1991, Taken on 13th May 2017, http://tdc-www.harvard.edu/catalogs/bsc5.readme
7. W. Gliese, H. Jahreiss, *Gliese catalog of Nearby Stars*, AstronomischesRechen-Institut, Germany, 1991, V/70A ftp://cdsarc.u-strasbg.fr/cats/V/70A/ReadMe
8. A. Ng, The EM Algorithm, Stanford University, 2016, taken on 13th May 2016, http://cs229.stanford.edu/notes/cs229-notes8.pdf
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 http://weka.sourceforge.net/doc.dev/weka/clusterers/EM.html
10. G. Ledrew, *The Real Starry Sky*, Journal of the Royal Astronomical Society of Canada, Vol. 95, p.32, 2001, DOI: 2001JRASC..95...32L
11. N. M. Ball, R. J. Brunner, *Data Mining and Machine Learning in Astronomy,* Int. J. Mod. Phys. D 19, 1049, 2010, DOI:S0218271810017160

12. N. M. Ball, R. J. Brunner, A.D. Myers, D. Tcheng, *Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees,* The American Astronomical Society, The Astrophysical Journal, Volume 650, Number 1, 2006
13. S. Jing, W. DeSheng, L. GuangRui, *An Efficient Guide Stars Classification Algorithm via Support Vector Machines,* Intelligent Computation Technology and Automation, 2009, DOI: 10.1109/ICICTA.2009.44
14. L. F. Smith, *A Revised Spectral Classification System and a New Catalogue for Galactic Wolf-Rayet Stars,* Mon Not R Astron Soc, 1968, DOI: 138.1.109
15. R. Humphreys, J. Martin, & M. Gordon, *Stellar Spectral Subclasses Classification Based on Fisher Criterion and Manifold Learning*, A New Luminous Blue Variable in M31, Publications of the Astronomical Society of the Pacific, 127(954), 789-794, 2015 DOI:10.1086/680998
16. J. Bloom, J. Richards, P. Nugent, et.al., *Automating Discovery and Classification of Transients and Variable Stars in the Synoptic Survey Era*, Publications of the Astronomical Society of the Pacific, 124(921), 1175-1196, 2012 doi:10.1086/668468