International Journal of Advanced Astronomy, 7 (1) (2019) 11-14

International Journal of Advanced Astronomy

Website: www.sciencepubco.com/index.php/IJAA

Research paper

Big data in astronomy: from evolution to revolution

M. Khalil [1] *, M. Said [1], H. Osman [1], B. Ahmed [2], D. Ahmed [3], N. Younis [4], B. Maher [3], M. Osama [3], M. Ashmawy [3]

[1] Department of mathematics, Faculty of engineering, Modern sciences and arts University (MSA), Egypt
[2] Department of electrical and computer engineering, Faculty of engineering, University of Victoria, Canada
[3] Department of electrical systems engineering, Faculty of engineering, Modern sciences and arts University (MSA), Egypt
[4] Department of English and American studies, University of Vienna, Austria
**Corresponding author E-mail: mkibrahim@msa.eun.eg*

Abstract

Big data is pushing astronomy in a new direction. Massive amounts of astronomical big data produced by the new generations of powerful instruments and simulations are exponentially gathered every day. Big data like astronomical images, infrared, microwave, ultraviolet, X-rays and gamma generated by stars, galaxies and black holes are observed by the new generations of space telescopes. It may take years to uncover the hidden signals in such data that may already hold answers to some of the fundamental questions of the universe we're seeking. In this paper, we attempt to present a short review about the astronomical big data and how can such massive data change our understanding of the universe.

Keywords: *Astronomical Big Data; Computational Mathematics & Statistics; Space Probes & Super Telescopes-Radio Telescopes; Machine Learning-Data Mining.*

1. Introduction

Astronomy is one of the oldest sciences in the history of humanity. During the last few decades, the rise of big data has taken over astronomy. Massive amount of data about the galaxies, stars, planets, comets, asteroids collected by space telescopes, satellites, and space probes are transferred to the Earth to be analyzed by scientists [1,2]. For instance, the "Sloan Digital Sky Survey"(SDSS) telescope produces 200 GB of data every night. Millions of field images have been acquired to detect millions of galaxies and stars [3]. Every 20 seconds, the 3200 megapixels camera of the "large synoptic survey telescope" (LSST) with its 3-ton digital camera captures one 6-Gigabyte image to detect 37 billion stars and galaxies recording the entire visible sky twice a week to produce 20 trillion bytes of raw data every night [4]. On the other hand, digitizing the old largest scanned astronomical plate images stored in 414 archives, comprising more than 2 million plates obtained by old telescopes to be accessible for scientists worldwide produce large scale-data [5].
Scientists define the big data characteristics as the big data 10 V's: Volume, Variety, Velocity, Veracity, Validity, Value, Variability, Venue, Vocabulary, and Vagueness [6,7]. The enormous big data forms like images, spectra, time series, functional data and radio waves give a better understanding of the Earth and universe evolution. Also such massive data provides scientists with the needed information in order to protect Earth from threats like comets and asteroids impacts [8]. Also such massive data allows us to estimate the age of galactic structures. The majority of the astronomical big data is based on astrophysical phenomena which are observed in terms of light snapshots which may not be repeated again. So, it is understood that why every signal observation is significant. Scientists have no clear idea about how to process the tidal wave of big data. Also, there is no enough servers to store such big data. Also there is not enough electricity to operate the servers as well. Automation offers the solutions of such problems [9]. With automated data processing, we are not time-limited. For instant, the Square Kilometer Array (SKA) which is a large multi radio telescope project in Australia and South Africa will be fully operated in 2020 to gather big data on the location and properties of stars, galaxies and clouds of hydrogen gas. Every second, SKA shrinks 5TB of astronomical data to 22MB in order to handle the big data [10].
The rest of the paper is organized as follows. In section 2, a brief discussion about the role of space probes in gathering astronomical big data is presented, while in section 3, the role of the sophisticated telescopes in the era of big data is explained. Section 4 is devoted for the impact of computational mathematics and statistics in studying and analyzing astronomical big data. We present the ideas of big data mining and machine learning in astronomy in section 5 and 6 respectively. The challenges and the future of big data in astronomy are presented briefly in section 7 followed by a conclusion in section 8.

2. Space probes used to collect astronomical big data

Beside the super telescopes that are used to collect the big data, space probes are significant tools used to gather big data. A space probe is an unmanned device, unpiloted spacecraft used to explore space, to gather big astronomical data from celestial bodies and other planet and to study the atmosphere. It is a robotic spacecraft that does not orbit the Earth, but, instead, explores further into outer space. A space probe is launched mainly from Earth with a set of scientific tools and instruments used to explore space. Rosetta which has been launched to

collect data from the comet 67P/Churyumov-Gerasimenko, New Horizons that has been launched to collect data from Pluto and Gaia that has been surveying our Milky Way [11] are considered famous examples of space probes. Most space probes transmit massive amounts of data from space by radio to Earth. The gathered data by space probes are huge data. Gaia collects massive data from about 1 billion stars and other bodies. One percent of such bodies are in the Milky Way galaxy [11]. New Horizons task is to observe at least two-dozen other Kuiper Belt objects (KBOs) beside the mission of collecting data from Pluto [12], while Rosetta target was to land on the comet Churyumov-Gerasimenko to study it and to collect the required data. Rosetta and its robotic lander Philae captured about 100,000 images during this mission. Such data reshaped our understanding of the comets [13].

3. The rise of advanced telescopes in the era of big data

About 400 years separate between Galileo's two-inch telescope and the huge nowadays sophisticated telescopes. The role of modern telescopes in gathering astronomical big data is significant. Super telescopes like Hubble space telescope and the LSST telescope which changed our understanding of the Universe [14]. They helped scientists to discover new planets, to understand the nature of Dark Matter, solar system, galaxies and stars. Radio telescopes in recent times have evolved rapidly. Radio telescopes provide us with answers of several questions in the field of astrophysics. Such telescopes provide us with better understanding of the nature of black holes and the evaluation of galaxies at radio wavelengths [15]. The total collecting area of the largest radio telescope SKA will be well over one 1 square kilometre [16-17]. "The SKA is going to be revolutionary compared with current telescopes because it's going to be built in such a way that we can extract even more information out of the data we receive from it," says Dr Cathyrn Trott, a Senior Research Fellow at the International Centre for Radio astronomy Research and member of SKA science and engineering committee [17].

4. Computational mathematics and statistics behind astronomical big data: an incomplete survey

In this section, we are trying to explain how computational mathematics and statistics play significant roles in astronomical big data science. For example, in optimization field, when it is needed to minimize a cost function which is based on the massive data, mathematical algorithms are required to solve such a problem. Distributed optimization enables us to apply several parallel optimization techniques at different computers at the same time to solve the mathematical optimization problems [18].
Statistics and data mining are crucial tools to understand the hidden knowledge behind the astronomical big data, we can start with a scalar quantity, x_i that is measured N times where $i = 1,2,3, \ldots N$ and the random variable X is the set of all outcomes x_i [19]. One of the big problems in astronomical big data mining is how to estimate the distribution h(x) from which values of x are drawn where h(x) expresses the probability that a value lies between x and $x + dx$, equal to h(x) dx, which is called a probability density function (pdf). The cumulative distribution function (cdf) can be written as

$$H(x) = \int_{-\infty}^{x} h(x')\, dx'$$

To identify the true pdf h(x) from a big data empirical pdf we recall the function $f(x)$ and its cumulative function $F(x)$. As it is assumed both $h(x)$ and $f(x)$ are normalized probability density functions so,

$$H(\infty) = \int_{-\infty}^{\infty} h(x')\, dx' = 1$$

And similarly, $F(\infty) = 1$. Since the data sets are not infinitely big, $f(x)$ cannot be exactly equal to $h(x)$. If we assume that, the measurement errors for x are not trivial then $f(x)$ will not approaches $h(x)$ even for infinitely big samples. $f(x)$ can be considered as a model of the true distribution $h(x)$; given that, only samples from $h(x)$ are observed. The functional forms (parametric or nonparametric models) of $h(x)$ have to be implemented to test it against the data to decide that if it is accepted or rejected.

The quantity x may be measured with some error distribution, $e(x)$, which is defined as the probability of measuring value x if the true value is μ,

$$e(x) = p(x|\mu, I)$$

where I is based on the information and the details of the error distribution. For example, the Gaussian error distribution where I is the standard deviation σ:

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-(x-\mu)^2}{2\,\sigma^2}\right)$$

Massive data sets require automated image processing and quality assessment of the processed images. Using partial differential equations (PDEs) in denoising and smoothing images is very effective [19]. So, the field of PDEs is a good example to explain how advanced mathematics can inspire the big data field. Recently, PDEs are used to tackle significant problems in big data science and to develop this growing field as well. PDEs have an essential role to find answers to astronomical big data questions. In image processing of astronomical big data images, solving an initial value Cauchy problem for the diffusion PDE with the noisy image [20] leads to filtering based on multi-scale representations of image data by the nonlinear PDE [21-23]. Such filtering helps in astronomical image enhancement and object detection. Perona & Malik [24] proposed the following nonlinear diffusion PDE whose diffusion coefficient D decreases when the gradient grows and increases when the gradient decays

$$u_t = div\,[D\,(\|\nabla u\|^2)\nabla u]$$

$$u(0, x, y) = u_0(x, y)$$

where $u(t,x,y)$ is the evolved image at the time t, and $u_0(x,y)$ is the noisy image [21, 25]. ∇u gives information about local intensity variations while the coefficient of diffusion D is defined as

$$D\ (\|\nabla u\|^2) = exp(-\|\nabla u\|^2\ k^{-2})$$

Where k presents the gradient scale of the initial image.
Statistics is an essential in big data analysis. The main goal of using statistics in big data science is to analyze the sample in order to estimate the population. It is impossible to analyze the entire big data due to its huge volume. We identify the big data sample set using advanced computing systems such as cloud computing. One of the strategies to handle and to analyze big data is to shrink it by identifying a subset of the entire data which keeps its mathematical relationships. One of the shrinking protocols is to estimate regression coefficients. For example, consider the regression model [26]:

$$Y = X_n\beta + \varepsilon$$

Where $Y = (y_1, y_2, \dots, y_n)^T$ is a vector of responses, X_n is an $n \times p$ matrix, $\beta = (\beta_1, \beta_2, \dots, \beta_n)^T$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$ is the vector of unobservable random errors.
Regression is a form of supervised machine learning as it helps to predict new data based on some existing data.

5. Astronomical big data mining

Big data mining is an essential tool that can be used to analyze big data in astronomy. It can help scientists to extract and discover useful information from the stored big data [6]. Data mining tasks are varying between summarization, classification, regression, clustering, association, time-series analysis, and outlier/anomaly detection [4]. Several data mining tools and software have been developed in the last few years in order to perform the data mining tasks and to overcome astronomical problems.

6. Machine learning methods to analyze astronomical big data

In astronomy, the new generations of telescopes such as Atacama Large Millimeter Array (ALMA) and Jansky VLA can provide vast amounts of data through large surveys (e.g., SDSS, ZTF, Pan-STARRS, VLT Survey Telescope (VST), ...etc.) cannot be analyzed manually. Also, new telescopes such as the Large Synoptic Survey Telescope (LSST) and the Square Kilometer Array (SKA) will become operational within the current decade and data volume will be even much increased. As data volume increases, analysis becomes more complicated and difficult to exploit for knowledge extraction. Therefore, it is important to develop new techniques for processing the amount and variety of astronomical big data to be capable of answering scientific questions based on the data. Machine learning is among those techniques that can be used to find the relation between input data (e.g., galaxy images) and outputs (e.g., physical properties of galaxies). For example, distances from Earth to galaxies, relative velocities of receding galaxies, and chemical compositions are measured by their redshifts which can be done using machine learning [27]. Since 2009, NASA's Kepler spacecraft has been searching for new planets known as exoplanet outside our solar system. Kepler is observing light curves of stars using automated software in order to detect changes in the stars brightness. These changes indicate whether a planet has moved in front of the star. If the brightness of the star appears to change with regular period, duration and decrease in brightness, this is likely due to an exoplanet. Any decrease in the brightness can be close to the noise level which makes it difficult to be detected by the traditional software. Vanderburg, a Nasa Segan fellow at UT Austin, and Shallue, a Google machine learning researches, presented a new method for classifying potential planet signals using deep learning [28]. They were able to statistically validate two new planets that are identified with high confidence using their model. Machine learning techniques cannot substitute physical models, as they do not supply scientific facts beyond the predicted data. In other words, machine learning and physical models support each other.

7. Challenges of the future of big data in astronomy

Big data are playing a significant role in the future of universe exploration. By using modern software and sophisticated hardware tools which help in analyzing and mining big data, universe exploration can be developed rapidly. Unfortunately, there are several challenges that face big data software and hardware. One of such challenges is hardware infrastructure, as big data analysis needs a complicated software and sometime we may lose big data due to hardware failure or software errors [29]. When big data are unstructured and heterogeneous it becomes more difficult to manage or to analyze such massive data using traditional software tools [30]. The computational requirements of big data analysis is bigger than the offered traditional computing resources [31].
Because of the limited available space for big data storage, it is necessary to identify the useful data and delete the rest quickly. So, we believe that scientists should develop their computer codes to manage the received big data. On the other hand, the scalability and huge data size are challenges that face machine learning techniques as they slow down the processing time [31]. Within the next few years, Astronomy will not only be multi-wavelength, but also multi-messenger, and dominated by massive data [32]. The size of astronomical big data will be increased rapidly over the next decade and will approach Petabyte scales. So, a commercial big data analytics processes should be considered. We argue that, advanced image processing and machine learning techniques can manage terabytes of massive data with high accuracy in near real-time during the next decade [3,37].

8. Conclusion

Big data in astronomy will change the way that we study astronomy. The massive data received every day motivate scientists to develop their sophisticated electronic instruments to manage and analyze such data. In this work, we argue that machine learning and data mining can offer several solutions to the big data problems. We hope to see a real integration among multiple scientific disciplines researches to overcome the obstacles and challenges that face the big data in astronomy.

References

[1] Yang, C., Huang, Q., Li, Z., Liu, K. and Hu, F., 2017. Big Data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth, 10(1), pp.13-53. https://doi.org/10.1080/17538947.2016.1239771.

[2] Feigelson, E.D. and Babu, G.J., 2012. Big data in astronomy. Significance, 9(4), pp.22-25. https://doi.org/10.1111/j.1740-9713.2012.00587.x.

[3] Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K.S. and Igel, C., 2017. Big universe, big data: machine learning and image analysis for astronomy. arXiv preprint arXiv:1704.04650. https://doi.org/10.1109/MIS.2017.40.

[4] LSST large telescope: https://www.lsst.org/

[5] Tsvetkov, M., 2005. Wide-field plate database: a decade of development. In Virtual Observatory: Plate Content Digitization, Archive Mining and Image Sequence Processing, iAstro workshop, Sofia, Bulgaria (pp. 10-41).

[6] Zhang, Y. and Zhao, Y., 2015. Astronomy in the big data era. Data Science Journal, 14. https://doi.org/10.5334/dsj-2015-011.

[7] Top 10 Big Data Challenges – A Serious Look at 10 Big Data V's: https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/.

[8] Long, J.P. and de Souza, R.S., 2017. Statistical methods in astronomy. arXiv preprint arXiv:1707.05834.

[9] Automation offers big solution to big data in astronomy: https://phys.org/news/2015-03-automation-big-solution-astronomy.html

[10] Square Kilometre Array precursor shrinks 5TB of data to 22MB – every second! https://www.theregister.co.uk/2017/01/18/murchison_radiotelescope_opens_the_science_firehose/.

[11] Space probes: http://www.astronoo.com/en/space-probes.html.

[12] NASA-New Frontiers: https://web.archive.org/web/20150415224640/http://discoverynewfrontiers.nasa.gov/missions/missions_nh.cfml.

[13] New Atlas-ESA completes massive archive of Rosetta images and data: https://newatlas.com/esa-rosetta-philae-image-archive/55147/.

[14] Universe today: https://www.universetoday.com/135506/rise-super-telescopes-build/.

[15] Norris, R.P., 2017. Extragalactic radio continuum surveys and the transformation of radio astronomy. Nature Astronomy, 1(10), p.671. https://doi.org/10.1038/s41550-017-0233-y.

[16] The SKA Project: https://www.skatelescope.org/the-ska-project/.

[17] New Atlas-The Square Kilometre Array: How the world's biggest telescope will revolutionize astronomy: https://newatlas.com/square-kilometre-array/53498/.

[18] Mathematics for big data: https://sites.google.com/view/mathbigdata/home.

[19] Ivezić, Ž., Connolly, A.J., VanderPlas, J.T. and Gray, A., 2014. Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton University Press. https://doi.org/10.23943/princeton/9780691151687.001.0001.

[20] Witkin, A. 1983, Proc. Int. Joint Conf. Artif. Intell., 1019.

[21] Pesenson, M., Roby, W. and McCollum, B., 2008. Multiscale astronomical image processing based on nonlinear partial differential equations. The Astrophysical Journal, 683(1), p.566. https://doi.org/10.1086/589276.

[22] Starck, J.-L., Murtagh, F., & Bijaoui, A. 1998, Image Processing and Data Analysis: The Multiscale Approach (Cambridge: Cambridge Univ. Press) https://doi.org/10.1017/CBO9780511564352.

[23] Starck, J.-L., & Murtagh, F. 2002, Astronomical Image and Data Analysis (Berlin: Springer) https://doi.org/10.1007/978-3-662-04906-8.

[24] Perona, P., & Malik, J. 1987, IEEE Trans. Pattern Analysis and Mach. Intell., 12, 629. https://doi.org/10.1109/34.56205.

[25] Pesenson, M.Z., Pesenson, I.Z. and McCollum, B., 2010. The data big bang and the expanding digital universe: High-dimensional, complex and massive data sets in an inflationary epoch. Advances in Astronomy, 2010. https://doi.org/10.1155/2010/350891.

[26] Yuzbasi, B.A.H.A.D.I.R., Arashi, M.O.H.A.M.M.A.D. and Ahmed, S.E., 2017. Big Data Analysis Using Shrinkage Strategies. arXiv preprint arXiv:1704.05074.

[27] A. A. Collister and O. Lahav. "ANNz: estimating photometric redshifts using artificial neural networks." Publications of the Astronomical Society of the Pacific 116, no.818, 2004. https://doi.org/10.1086/383254.

[28] C. J. Shallue, and A. Vanderburg. "Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90." The Astronomical Journal 155, no.2, 2018. https://doi.org/10.3847/1538-3881/aa9e09.

[29] Adrian, A. (2013). Big Data Challenges. Database Systems Journal, 4 (3), 31-40.

[30] Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu & Fei Hu (2017) Big Data and cloud computing: innovation opportunities and challenges, International Journal of Digital Earth, 10:1, 13-53, https://doi.org/10.1080/17538947.2016.1239771.

[31] Wadhwani, K. (2017). Big Data Challenges and Solution..

[32] Garrett, M.A., 2014. Big Data analytics and cognitive computing–future opportunities for astronomical research. In IOP Conference Series: Materials Science and Engineering (Vol. 67, No. 1, p. 012017). IOP publishing https://doi.org/10.1088/1757-899X/67/1/012017.

[33] Edwards, K., Gaber, M.M.: Astronomy and Big Data: A Data Clustering Approach to Identifying Uncertain Galaxy Morphology, 1st edn. Springer, Heidelberg (2014).