Analysis of Exploratory Data Analysis Tools and Techniques

AVADHESH KUMAR GUPTA
Unitedworld School of Computational Intelligence , Karnavati University (Gujarat)- INDIA, (e-mail:: dr.avadheshgupta@gmail.com)

ABSTRACT The practice of doing exploratory analyses on large data sets is commonplace in several scientific disciplines. It is dependent on sophisticated data processing pipelines and computational models, which are often developed via interdisciplinary collaboration by scientists familiar with a wide range of programming languages, databases, and computing platforms. The tremendous complexity of both the data and the implemented analysis procedures, as well as the continual reuse and adaption of data analysis pipelines in multiple application scenarios, create significant ambiguity regarding the veracity of analysis outputs. In this article, we analyze the development in the field of exploratory data analysis with the help of the Scopus database.

KEYWORDS Exploratory Data Analysis; Big data; Scopus

I. INTRODUCTION

Data science is now ubiquitous and has an effect on almost every industry. Data is now essential to every field and business. Therefore, enterprises may now draw power from data science. Data science is an interdisciplinary subject that uses methods from other fields to gather data, analyze it, get new insights from it, and apply those insights to decision-making. Data science is a broad study that incorporates several areas of computer science, such as data mining, statistics, machine learning, data analytics, and programming languages like Python and R. Data science end-to-end solutions are built using a systematic approach that spans the whole process, from defining the issue and gathering relevant data through testing and releasing the final product. Performing analysis of the data is known as exploratory data analysis (EDA), a step that happens at the beginning of the data science life cycle. After using EDA methods, individuals are more attuned to the source data and more aware of any abnormalities or discrepancies [1], [2]. It might be time-consuming and arduous to manually evaluate the data and apply all of the EDA approaches. In this context, we analyze the development in the field of EDA.

II. LITERATURE SURVEY

Author in [3] proposed a DDoS Attacks [4] and Defense Mechanisms in Various Web-Enabled Computing Platforms: Issues, Challenges, and Future Research Directions. Author in [5] proposed a secure and energy efficient-based E-health care framework for green internet of things. Author in [6] proposed a novel coverless information hiding method based on the average pixel value of the sub-images. Author in [7] proposed a secure Machine Learning scenario from Big Data in Cloud Computing via Internet of Things network. Author in [8] proposed a context Aware Recommender Systems. Author in [9] proposed a cross-lingual transfer method and distributed MinIE algorithm on apache spark. Author in [10] proposed a smart defense against distributed Denial of service attack in IoT networks using supervised learning classifiers. Author in [11] proposed an adaptive Feature Selection and Construction for Day-Ahead Load Forecasting using Deep Learning Method. Author in [12] proposed and analysis of artificial intelligence-based technologies and approaches on sustainable entrepreneurship. Author in [13] proposed a digital Watermarking-Based Cryptosystem for Cloud Resource Provisioning. Author in [14] presents a Browser-Side Context-Aware Sanitization of Suspicious HTML5 Code for Halting the DOM-Based XSS Vulnerabilities in Cloud. Author in [15] proposed a lightweight mutual authentication protocol based on elliptic curve cryptography for IoT devices. Author in [16] proposed a secure Timestamp-Based Mutual Authentication Protocol for IoT Devices Using RFID Tags. Author in [17] proposed a novel framework for risk assessment and resilience of critical infrastructure towards climate change. Author in [18] presets a review on advances in security and privacy of multimedia big data in mobile and cloud computing. Author in [19] proposed myocardial infarction detection based on deep neural network on imbalanced data. Author in [20] proposed a reputation score policy and Bayesian game theory based incentivized mechanism for DDoS attacks mitigation and cyber defense.

III. RESEARCH METHODOLOGY

In this article, we analyze the development in the field of Exploratory Data Analysis. We search the Scopus database

using the following query:

TITLE-ABS-KEY ("Exploratory Data Analysis") AND (LIMIT-TO (SUBJAREA , "COMP"))

The above-defined query extracts all the articles that have “Exploratory Data Analysis" in their title, abstract, or keywords.

IV. RESULT AND DISCUSSION

In this research, we analyze the work in the field of exploratory data analysis. As explained in the previous section, we used the Scopus database to conduct our research. After running the query, we get 1912 documents as represented in Figure 1. Figure 2 presents the number of papers published over the time-span related to exploratory data analysis. From Figure 2 it is clear that the average growth rate of papers over that year is 5.37%, this shows that exploratory data analysis is a relevant topic and needs further research. The collected documents are published at different platforms as represented in Figure 3. however, it is clear that the majority of the articles are published in international conferences (53.5%). In addition to that, from Figure 4 it is clear that the majority of computer science researchers are working in the field of exploratory data analysis. In the next subsections, we analyze the Scopus database with different parameters.

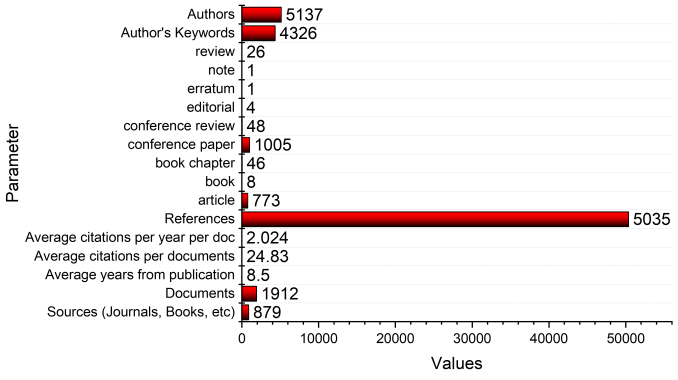

FIGURE 1: General Information

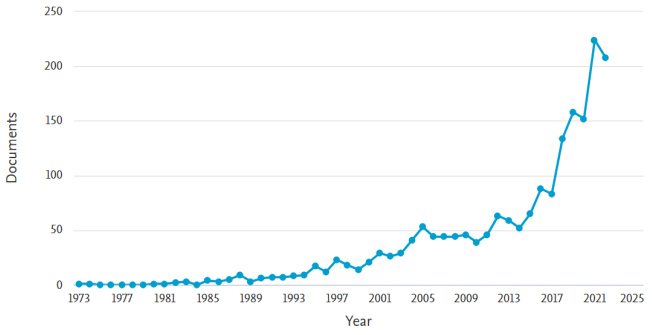

FIGURE 2: Annual Production

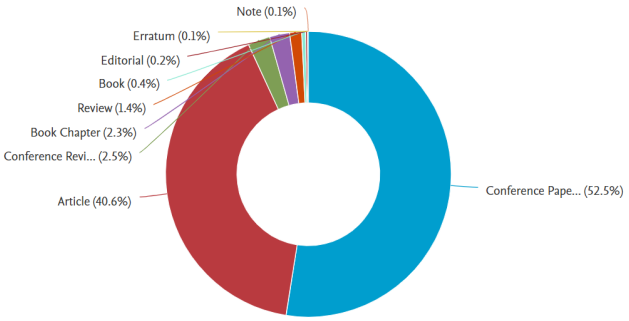

FIGURE 3: Type of Publication

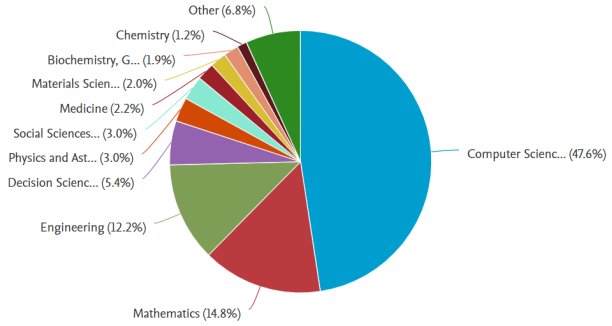

FIGURE 4: Different Domain Information

A. ANALYSIS OF AUTHORS

In this subsection, we analyze the author distribution. We arrange the authors according to the number of published papers. This arrangement is represented as follows:

- NA NA (50)
- LANGE O (11)
- WU Y (10)
- ANDRIENKO N (9)
- COOK D (9)
- FYFE C (9)
- SHNEIDERMAN B (9)
- ANDRIENKO G (8)
- KASKI S (8)
- TAKAMA Y (8)
- WISMÜLLER A (8)

This subsection helps us to find the leading researcher in the field of exploratory data analysis. This will help young researchers to find relevant articles on exploratory data analysis.

B. ANALYSIS OF DOCUMENT DISTRIBUTION

In this subsection, we detail the publications’ dissemination in the scientific community. Our study included 1912 works from Scopus-indexed periodicals, conferences, reviews, books, and book chapters. The most-cited publications in the subject of exploratory data analysis provide an overview of the many issues and significant ideas put out

TABLE 1: Highly Cited Papers

Paper	DOI	Total Citations
JAIN AK, 1999, ACM COMPUT SURV [21]	10.1145/331499.331504	9889
LÊ S, 2008, J STAT SOFTWARE [22]	10.18637/jss.v025.i01	4862
VESANTO J, 2000, IEEE TRANS NEURAL NETWORKS [23]	10.1109/72.846731	1896
BADDELEY A, 2005, J STAT SOFTWARE [24]	10.18637/jss.v012.i06	1636
FRIEDMAN JH, 1974, IEEE TRANS COMPUT [25]	10.1109/T-C.1974.224051	1179
KOHONEN T, 2000, IEEE TRANS NEURAL NETWORKS [26]	10.1109/72.846729	700
ZHANG T, 1997, DATA MIN KNOWL DISCOV [27]	10.1023/A:1009783824328	565
VIÉGAS FB, 2004, CONF HUM FACT COMPUT SYST PROC [28]	10.1145/985692.985765	557
MAO J, 1995, IEEE TRANS NEURAL NETWORKS [29]	10.1109/72.363467	524
ANDRIENKO N, 2006, EXPLORATORY ANAL OF SPAT AND TEMPORAL DATA: A SYST APPROACH [30]	10.1007/3-540-31190-4	454
RAO R, 1994, CONF HUM FACT COMPUT SYST PROC [31]	NA	441
RISSO D, 2011, BMC BIOINFORM [32]	10.1186/1471-2105-12-480	436
HECKERMAN D, 2001, J MACH LEARN RES [33]	NA	388
RAUBER A, 2002, IEEE TRANS NEURAL NETWORKS [34]	10.1109/TNN.2002.804221	375
HAMED MM, 2004, ENVIRON MODEL SOFTW [35]	10.1016/j.envsoft.2003.10.005	347
GE SX, 2018, BMC BIOINFORM [36]	10.1186/s12859-018-2486-6	345
BRECHMANN EC, 2013, J STAT SOFTWARE [37]	10.18637/jss.v052.i03	295
GÖNEN M, 2012, BIOINFORMATICS [38]	10.1093/bioinformatics/bts360	291
ANDRIENKO GL, 1999, INT J GEOGR INF SCI [39]	10.1080/136588199241247	268
HENDERSON K, 2012, PROC ACM SIGKDD INT CONF KNOWL DISCOV DATA MIN [40]	10.1145/2339530.2339723	264

by scholars. Table 1 presents the distribution of the paper according to the total number of citations.

C. ANALYSIS OF KEYWORDS

As we know, keywords give an overview of the research articles. Therefore analyzing the keyword distribution of the Scopus database gives a brief representation of the research domain. Figure 5 presents the keyword distribution. In the Figure 5, keywords are arranged according to the occurrence frequency; as the occurrence frequency of a keyword increases, its size increases. Therefore, from Figure 5 it is clear that frequently occurring keywords are as follows:

FIGURE 5: Keywords Distribution

- exploratory data analysis (531)
- machine learning (129)
- data mining (101)
- clustering (84)
- data visualization (67)
- visualization (60)
- visual analytics (49)
- data analysis (43)
- classification (42)
- dimensionality reduction (41)

D. ANALYSIS OF COUNTRY

The locations of researchers are likewise crucial to the advancement of scientific inquiry. This section thus provides an evaluation of the impact of location on research in EDA. Figure 6 presents the result of our analysis. We can compute the ranking of the countries according to the number of papers their researchers published. From Figure 6; the ranking of the countries are as follows.

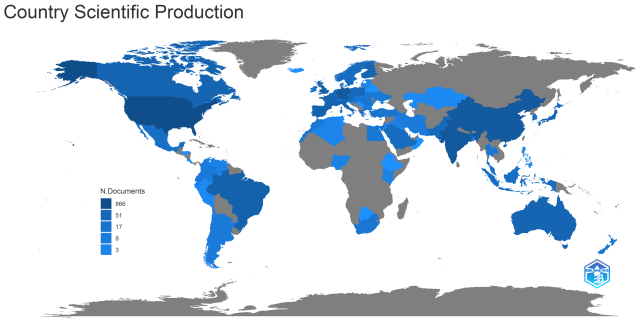

FIGURE 6: Country Production

- USA (866)
- INDIA (302)
- GERMANY (247)
- CHINA (234)
- UK (163)

- SPAIN (120)
- BRAZIL (114)
- FRANCE (89)
- ITALY (83)
- CANADA (80)

V. CONCLUSION

Exploratory Data Analysis (EDA) is a method of data analysis that depends on the use of graphical tools. It may be used to find patterns and trends, as well as to check assumptions and hypotheses, using statistical summaries and visual representations. Not only can it help discover glaring mistakes, but it can also shed light on hidden meanings, flag out-of-the-ordinary occurrences, unearth surprising connections between variables, and more.

REFERENCES

[1] R. K. S. Rajput, D. Goyal, A. Pant, G. Sharma, V. Arya, and M. K. Rafsanjani, "Cloud data centre energy utilization estimation: Simulation and modelling with idr," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–16, 2022.

[2] K. Pathoee, D. Rawat, A. Mishra, V. Arya, M. K. Rafsanjani, and A. K. Gupta, "A cloud-based predictive model for the detection of breast cancer," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–12, 2022.

[3] A. Singh and et al., "Distributed denial-of-service (ddos) attacks and defense mechanisms in various web-enabled computing platforms: Issues, challenges, and future research directions," International Journal on Semantic Web and Information Systems (IJSWIS), vol. 18, no. 1, pp. 1–43, 2022.

[4] A. Gaurav, V. Arya, and D. Santaniello, "Analysis of machine learning based ddos attack detection techniques in software defined network," Cyber Security Insights Magazine (CSIM), vol. 1, no. 1, pp. 1–6, 2022.

[5] M. Kaur and et al., "Secure and energy efficient-based e-health care framework for green internet of things," IEEE Transactions on Green Communications and Networking, vol. 5, no. 3, pp. 1223–1231, 2021.

[6] L. Zou and et al., "A novel coverless information hiding method based on the average pixel value of the sub-images," Multimedia tools and applications, vol. 78, no. 7, pp. 7965–7980, 2019.

[7] C. L. Stergiou and et al., "Secure machine learning scenario from big data in cloud computing via internet of things network," in Handbook of computer networks and cyber security. Springer, 2020, pp. 525–554.

[8] M. Casillo and et al., "Context aware recommender systems: A novel approach based on matrix factorization and contextual bias," Electronics, vol. 11, no. 7, p. 1003, 2022.

[9] P. Do and et al., "Building a knowledge graph by using cross-lingual transfer method and distributed minie algorithm on apache spark," Neural Computing and Applications, pp. 1–17, 2020.

[10] B. Gupta, P. Chaudhary, X. Chang, and N. Nedjah, "Smart defense against distributed denial of service attack in iot networks using supervised learning classifiers," Computers & Electrical Engineering, vol. 98, p. 107726, 2022.

[11] R. Jiao and et al., "Adaptive feature selection and construction for day-ahead load forecasting use deep learning method," IEEE Transactions on Network and Service Management, vol. 18, no. 4, pp. 4019–4029, 2021.

[12] B. B. Gupta, A. Gaurav, P. K. Panigrahi, and V. Arya, "Analysis of artificial intelligence-based technologies and approaches on sustainable entrepreneurship," Technological Forecasting and Social Change, vol. 186, p. 122152, 2023.

[13] S. Kumar, S. Kumar, N. Ranjan, S. Tiwari, T. R. Kumar, D. Goyal, G. Sharma, V. Arya, and M. K. Rafsanjani, "Digital watermarking-based cryptosystem for cloud resource provisioning," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–20, 2022.

[14] B. B. Gupta, S. Gupta, and P. Chaudhary, "Enhancing the browser-side context-aware sanitization of suspicious html5 code for halting the dom-based xss vulnerabilities in cloud," International Journal of Cloud Applications and Computing (IJCAC), vol. 7, no. 1, pp. 1–31, 2017.

[15] A. Tewari and et al., "A lightweight mutual authentication protocol based on elliptic curve cryptography for iot devices," International Journal of Advanced Intelligence Paradigms, vol. 9, no. 2-3, pp. 111–121, 2017.

[16] A. Tewari and et al., "Secure timestamp-based mutual authentication protocol for iot devices using rfid tags," International Journal on Semantic Web and Information Systems (IJSWIS), vol. 16, no. 3, pp. 20–34, 2020.

[17] N. Kumar and et al., "A novel framework for risk assessment and resilience of critical infrastructure towards climate change," Technological Forecasting and Social Change, vol. 165, p. 120532, 2021.

[18] B. B. Gupta, S. Yamaguchi, and D. P. Agrawal, "Advances in security and privacy of multimedia big data in mobile and cloud computing," Multimedia Tools and Applications, vol. 77, no. 7, pp. 9203–9208, 2018.

[19] M. Hammad and et al., "Myocardial infarction detection based on deep neural network on imbalanced data," Multimedia Systems, vol. 28, no. 4, pp. 1373–1385, 2022.

[20] A. Dahiya and et al., "A reputation score policy and bayesian game theory based incentivized mechanism for ddos attacks mitigation and cyber defense," Future Generation Computer Systems, vol. 117, pp. 193–204, 2021.

[21] A. Jain, M. Murty, and P. Flynn, "Data clustering: A review," ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[22] S. Lê, J. Josse, and F. Husson, "Factominer: An r package for multivariate analysis," Journal of Statistical Software, vol. 25, no. 1, pp. 1–18, 2008.

[23] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 586–600, 2000.

[24] A. Baddeley and R. Turner, "spatstat: An r package for analyzing spatial point patterns," Journal of Statistical Software, vol. 12, pp. 1–42, 2005.

[25] J. Friedman and J. Tukey, "A projection pursuit algorithm for exploratory data analysis," IEEE Transactions on Computers, vol. C-23, no. 9, pp. 881–890, 1974.

[26] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 574–585, 2000.

[27] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: A new data clustering algorithm and its applications," Data Mining and Knowledge Discovery, vol. 1, no. 2, pp. 141–182, 1997.

[28] F. Viégas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," 2004, pp. 575–582.

[29] J. Mao, "Artificial neural networks for feature extraction and multivariate data projection," IEEE Transactions on Neural Networks, vol. 6, no. 2, pp. 296–317, 1995.

[30] N. Andrienko and A. Gennady, Exploratory analysis of spatial and temporal data: A systematic approach, 2006.

[31] R. Rao and S. K. Card, "Table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information," 1994, pp. 318–322.

[32] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, "Gc-content normalization for rna-seq data," BMC Bioinformatics, vol. 12, no. 1, 2011.

[33] D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie, "Dependency networks for inference, collaborative filtering, and data visualization," Journal of Machine Learning Research, vol. 1, no. 1, pp. 49–75, 2001.

[34] A. Rauber, D. Merkl, and M. Dittenbach, "The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data," IEEE Transactions on Neural Networks, vol. 13, no. 6, pp. 1331–1341, 2002.

[35] M. Hamed, M. Khalafallah, and E. Hassanien, "Prediction of wastewater treatment plant performance using artificial neural networks," Environmental Modelling and Software, vol. 19, no. 10, pp. 919–928, 2004.

[36] S. Ge, E. Son, and R. Yao, "idep: An integrated web application for differential expression and pathway analysis of rna-seq data," BMC Bioinformatics, vol. 19, no. 1, 2018.

[37] E. Brechmann and U. Schepsmeier, "Modeling dependence with c- and d-vine copulas: The r package cdvine," Journal of Statistical Software, vol. 52, no. 3, pp. 1–27, 2013.

[38] M. Gönen, "Predicting drug-target interactions from chemical and genomic kernels using bayesian matrix factorization," Bioinformatics, vol. 28, no. 18, pp. 2304–2310, 2012.

[39] G. Andrienko and N. Andrienko, "Interactive maps for visual data exploration," International Journal of Geographical Information Science, vol. 13, no. 4, pp. 355–374, 1999.

[40] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "Rolx: Structural role extraction mining in large graphs," 2012, pp. 1231–1239.