Research Approach

The research approach adopted by Chen et al. (2024) is primarily experimental, focusing on the development and evaluation of the Data-Driven Multinomial Random Forest (DMRF) algorithm. The study involves systematic experimentation to compare the performance of DMRF with existing algorithms such as Multinomial Random Forest (MRF), Bernoulli Random Forest (BRF), and Denil14 (Poisson RF). The experimental design includes manipulating variables to explore the effects and conducting comparative analyses across different performance metrics (Quick & Hall, 2015). Additionally, the research utilizes datasets from the UCI database and employs techniques like bootstrapping and padding operation to enhance algorithm performance and handle missing values. This experimental approach allows for a comprehensive evaluation of the DMRF algorithm's effectiveness and efficiency compared to its counterparts, providing valuable insights into its potential applications and limitations.

Critical Analysis of Research Paradigm

Ontological Perspective

Based on the structure of the journal, it adopts a positivist ontological perspective (Park et al., 2020), as evidenced by several key elements:

1. *Nature of Reality*

The research conduct under the assumption of measurable truth, facilitating the objective observation and quantification of empirical reality. This is demonstrated through a comparative analysis aiming to quantify the evaluation of DMRF performance alongside existing algorithms in the field. The use of metrics such as accuracy, precision, and recall further supports this approach, indicating a focus on measurable and observable outcomes. Additionally, the research aligns with

the Hypothetico-Deductive Model, which forms the underlying principle of the positivist paradigm (Park et al., 2020).

2. *Quantitative Analysis*

The evaluation of the DMRF algorithms relies on empirical findings to facilitate comparison. The researcher employs hypothesis testing and regression analysis as statistical methods to assess the performance of the algorithms.

3. *Generalizability and Replicability*

According to Park et al. (2020), emphasizes research in controlled settings where variables can be manipulated. This approach values exploring relationships between dependent and independent variables, emphasizing replicability for validating findings from the literature review. Additionally, the journal assesses DMRF algorithm performance against baseline algorithms like MRF and BRF, highlighting differences from a statistical standpoint and ensuring generalizability of the research outcomes.

Epistemological Perspective

The journal adopts a pragmatic epistemology perspective, emphasizing the practical utility of knowledge (Ruwhiu & Cone, 2010; Johnson & Onwuegbuzie, 2004). This approach highlights the importance of applying research findings to real-world scenarios. The research conducted a comparative analysis to assess the practical effectiveness of the DMRF algorithm in various settings. Additionally, the study utilized standardized datasets to ensure relevance and applicability to practical contexts, aligning with the pragmatic emphasis on practical outcomes.

1. *Real World Application & Practical Implications*

The pragmatic approach showcased in the evaluation of the DMRF algorithm effectiveness in classification and regression tasks. Utilizing the UCI database ensures replicability throughout the

research process. This dataset mirrors real-case scenarios, enabling the evaluation of the DMRF algorithm from a practical perspective, thereby understanding the practical consequences and significance of the acquired knowledge.

2. *Empirical Evidence & Comparative Analysis*

By providing tangible evidence of the practical consequences, it serves as a method to validate the effectiveness of algorithms in addressing the practical problem. Therefore, the algorithms are evaluated statistically to determine the performance in real case scenarios.

Strength of Research Approach

1. *Measuring Variable*

The quantitative research approach is grounded in the positivist paradigm, prioritizing the objective measurement and analysis of variables. The journal conducts quantitative methods to evaluate the performance of the DMRF algorithm. Variables such as accuracy, precision, recall, and computational complexity are systematically measured and analyzed to assess DMRF effectiveness in classification and regression tasks (Rahman, 2017). By adhering to the positivist paradigm, the research aims to generate empirical evidence which is reliable and replicable.

2. *Comparison Analysis*

By evaluating both algorithm’s performance and the impact of dataset variations, the current research presents the strength of conducting a comparative analysis. By systematically comparing the DMRF algorithm between its existing counterparts, the researchers utilize quantitative metrics to gauge the relative effectiveness, as seen in Image 1. This approach facilitates an understanding of each algorithm's capabilities and limitations, supported by statistical methods (Castro et al., 2010). Additionally, the study delves into the influence of dataset size on algorithm performance.

Consistent findings by Althnian et al. (2021), the research suggests that larger datasets may compromise algorithm performance, evidenced by exponential growth in standard deviation. This aspect underscores the importance of dataset considerations in algorithm evaluation. Overall, the research methodology enables a systematic comparison of algorithms, incorporating analyses such as Standard Deviation Analysis, Computational Complexity Analysis, and Comparative Analysis between existing algorithms.

Datasets	DMRF	MRF(SE)	MRF(b)	BRF(SE)	BRF(b)	Denil14(SE)	Denil14(b)	BreimanRF
Blogger	**81.8***	76.2	79.9	78.3	81.2	75.8	80.5	81.4
Bone marrow	**93.96***	93.56	93.71	93.53	93.86	93.92	93.45	93.42
Algerian Forest Fires	**93.03***	92.25	92.16	92.71	93.71	92.45	92.46	93
Vertebral	**84.39**	83.19	83.58	83.74	84.35	82.74	82.74	84.64*
Chronic kidney disease	**98.80***	98.13	98.6	98.23	98.65	98.23	98.48	98.77
Cvr	**96.19***	95.49	95.61	95.91	96.16	95.63	95.74	95.84
House-votes	**96.17***	95.67	95.49	95.58	95.89	95.66	95.67	95.87
Wdbc	**96.25***	95.58	96.22	95.12	95.96	95.55	96.08	94.18
Breast original	**95.88**	95.29	95.48	95.48	95.72	94.45	94.69	96.71*
Balance scale	**83.45**	80.58	77.64	81.83	83.46	80.15	77.02	86.19*
Raisin	**86.22***	85.47	85.94	85.60	86.02	86.16	85.53	84.98
Vehicle	**75.63***	73.24	75.60	72.79	74.44	73.04	74.55	74.46
Tic-tac-toe	98.27*	98.47	**98.85**	98.18	98.08	97.82	98.38	94.37
HCV	**23.84**	23.55	23.66	23.25	23.28	23.28	23.22	24.88*
Winequality (red)	**69.83***	62.47	69.57	62.48	69.75	62.08	69.49	64.70
Wireless	**98.36***	98.28	98.33	98.2	98.18	97.78	97.97	98.33
Obesity	**78.54**	24.41	77.34	71.35	78.51	73.28	77.20	94.42*
Ad	97.66*	96.76	**97.95**	94.43	97.46	94.16	96.98	97.02
Spambase	**95.18***	93.6	95.01	93.93	95.02	91.48	95.1	91.82
Winequality (white)	**69.21***	59.93	69.20	60.65	69.44	60.07	68.71	64.02
Page blocks	**97.59***	97.44	97.56	97.17	97.45	97.28	97.44	97.06
MFCCs	**98.53***	98.02	98.5	98.02	98.47	97.83	98.36	63.57
Mushroom	57.24*	59.98	47.42	**62.10**	58.67	58.99	48.54	47.28
Ai4i	**59.96***	59.5	57.01	59.4	59.95	59.6	56.66	56.15
Letter	**89.79**	89.55	89.58	83.05	89.00	81.78	87.50	96.32*
Adult	**86.45***	86.28	86.13	57.57	86.44	86.19	85.57	85.98
Connect-4	82.18*	81.96	**84.08**	78.86	80.77	81.28	83.44	81.46

Image 1 (Comparison of accuracy in different model)

Weakness of Research Approach

1. *Lack of Domain Expertise*

The weakness lies in the lack of domain-specific expertise and the broad scope of application. By utilizing the UCI dataset without domain-specific insights, the research risks overlooking crucial nuances and requirements specific to industries, potentially resulting in algorithms that are not optimized for practical use. Moreover, the application of the model across diverse datasets without adjusting it to specific industries may lead to suboptimal performance and limited applicability. For example, the medical sector's unique precision and error margin requirements, which may not be addressed. Consequently, the research's failure to focus on industry-specific needs could prevent its practical relevance and adoption.

2. *Emphasize of hypothesis testing over hypothesis generation*

The research approach in the journal exhibits a tendency towards hypothesis testing rather than hypothesis generation, aligning with the identified weakness. Throughout the study, there is a clear emphasis on testing the performance of the DMRF algorithm and comparing it with existing algorithms, rather than exploring new theories or hypotheses. This focus on hypothesis testing may inadvertently lead to confirmation bias, as researchers may prioritize confirming existing hypotheses rather than considering alternative explanations or exploring new ideas (Johnson & Onwuegbuzie, 2004). Consequently, the research may miss out on valuable insights or alternative perspectives that could enrich the understanding of the algorithm's effectiveness and limitations. Therefore, while hypothesis testing is valuable for validating hypotheses and assessing algorithm performance, an overemphasis on this approach may restrict the exploration of new theories or phenomena within the research domain, as observed in the journal.

Recommendations for Future Research

1. *Industry-Specific Analysis via qualitative methods.*

Future research should prioritize industry-specific analysis to enhance the relevance and applicability of algorithms in real-world settings. By focusing on specific industries such as healthcare, finance, or manufacturing, researchers can gain deeper insights into the unique requirements and challenges of each sector. This approach enables the development of customise the algorithms to address the specific needs of different industries, leading to more effective and practical solutions. Additionally, industry-specific analysis allows for a more thorough evaluation of algorithm performance in relevant contexts, ensuring that research outcomes translate into tangible benefits for end-users. Therefore, incorporating industry-specific analysis into research methodologies can significantly enhance the impact and utility of algorithmic solutions across various domains (Johnson & Onwuegbuzie, 2004).

2. *Enhancing Research Depth.*

Integrating hypothesis generation into the research, alongside hypothesis testing, can enhance its depth and breadth. Researchers can begin with exploratory studies or qualitative research methods to develop new theories or hypotheses regarding the DMRF algorithm's performance. This may involve interviewing industry experts, stakeholders, or end-users to understand their perspectives and challenges. By exploring diverse viewpoints, researchers can formulate hypotheses that complement the existing framework (Farris & Farris, 1989; Ivankova & Wingo, 2018). These hypotheses can then be tested using quantitative methods to validate their relevance. This integrated approach enriches research findings and provides an understanding of the algorithm's effectiveness across industries. Ultimately, combining hypothesis generation with testing strengthens the research methodology and leads to more robust outcomes.

References

1. Althnian, A., AlSaeed, D., Al-Baity, H. H., Samha, A. K., Dris, A. B., Alzakari, N., … & Kurdi, H. (2021). Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Applied Sciences, 11(2), 796. https://doi.org/10.3390/app11020796
2. Castro, F. G., Kellison, J. G., Boyd, S. J., & Kopak, A. M. (2010). A methodology for conducting integrative mixed methods research and data analyses. Journal of Mixed Methods Research, 4(4), 342-360. https://doi.org/10.1177/1558689810382916
3. Chen, J. H., Wang, X. L., & Lei, F. (2024). Data-driven multinomial random forest: a new random forest variant with strong consistency. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-023-00874-6
4. Farris, H. H., & Revlin, R. (1989). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. Memory & Cognition, 17(2), 221–232. https://doi.org/10.3758/bf03197071
5. Ivankova, N., & Wingo, N. (2018). Applying Mixed Methods in Action Research: Methodological Potentials and Advantages. American Behavioral Scientist, 62(7), 978–997. https://doi.org/10.1177/0002764218772673
6. Johnson, R. B. and Onwuegbuzie, A. J. (2004). Mixed methods research: a research paradigm whose time has come. Educational Researcher, 33(7), 14-26. https://doi.org/10.3102/0013189x033007014
7. Park, Y. S., Konge, L., & Artino, A. R. (2020). The Positivism Paradigm of Research. Academic Medicine, 95(5), 690–694. https://doi.org/10.1097/ACM.0000000000003093

8. Quick, J., & Hall, S. (n.d.). *Part Three: The Quantitative Approach*. https://doi.org/https://doi.org/10.1177/175045891502501002
9. Rahman, S. (2017). The Advantages and Disadvantages of Using Qualitative and Quantitative Approaches and Methods in Language “Testing and Assessment” Research: a Literature Review. Journal of Education and Learning, 6(1), 102–112.
10. Ruwhiu, D., & Cone, M. (2010). Advancing a pragmatist epistemology in organisational research. *Qualitative Research in Organizations and Management: An International Journal*, *5*(2), 108–126. https://doi.org/10.1108/17465641011068884

Appendix

Chen et al. Journal of Big Data (2024) 11:34
https://doi.org/10.1186/s40537-023-00874-6

Journal of Big Data

METHODOLOGY Open Access

Data-driven multinomial random forest: a new random forest variant with strong consistency

JunHao Chen[1*], XueLi Wang[1*] and Fei Lei[2]

*Correspondence:
chenjunhao@emails.bjut.edu.cn;
xlwang@bjut.edu.cn

[1] School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing, China
[2] Faculty of Information Technology, Beijing University of Technology, Beijing, China

Abstract

In this paper, we modify the proof methods of some previously weakly consistent variants of random forest into strongly consistent proof methods, and improve the data utilization of these variants in order to obtain better theoretical properties and experimental performance. In addition, we propose the Data-driven Multinomial Random Forest (DMRF) algorithm, which has the same complexity with BreimanRF (proposed by Breiman) while satisfying strong consistency with probability 1. It has better performance in classification and regression tasks than previous RF variants that only satisfy weak consistency, and in most cases even surpasses BreimanRF in classification tasks. To the best of our knowledge, DMRF is currently a low-complexity and high-performing variation of random forest that achieves strong consistency with probability 1.

Keywords: Random forest, Strong consistency, Classification, Regression, Machine learning

Introduction

Random Forest (RF, also called standard RF or BreimanRF) [1] is an ensemble learning algorithm that makes classification or regression predictions by taking the majority vote or average of the results of multiple decision trees. Due to its simple and easy-to-understand nature, rapid training, and good performance, it is widely used in many fields, such as data mining [2–4], computer vision [5–7], ecology [8, 9], and bioinformatics [10].

Although the RF has excellent performance in practical problems, analyzing its theoretical properties is quite difficult due to its highly data-dependent tree-building process. These theoretical properties include consistency, which can be weak or strong. Weak consistency refers to the expectation of the algorithm's loss function converges to the minimum value as the data size tends to infinity, while strong consistency refers to the algorithm's loss function converges to the minimum value as the data size tends to infinity [11]. Consistency is an important criterion for evaluating whether an algorithm is excellent, especially in the era of big data.

Many researchers have made important contributions to the discussion of consistency-related issues in RF, proposing many variants of RF with weak consistency, such