Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

An Exploration of Parallel Data Mining Algorithm Implementation with Advanced Data Mining to Enhance Scalability

R.Rathiga[1]

Research Scholar

Dr.T.Rathimala[2] *Programmer/Assistant Professor*

Department of Computer and Information Science, Faculty of Science

Annamalai University.

Abstract

This research investigates the transformative impact of data mining techniques on data scalability, emphasizing the efficiency and adaptability of advanced algorithms. The study explores how these techniques optimize processing efficiency, allowing for streamlined analyses and reducing the computational burden associated with vast datasets. Selective resource utilization ensures a focused approach, enhancing the scalability of data mining processes tailored to specific analytical goals. Parallel Data Mining Algorithms, designed to distribute computational workloads across multiple processors, prove instrumental in overcoming challenges posed by large-scale datasets. The research also delves into the sophistication of advanced algorithms such as Random Forest and Neural Networks, providing versatile solutions for pattern recognition and insight extraction. Findings demonstrate the holistic contribution of these methodologies to scalable data analysis in diverse fields. This study offers insights for businesses and industries navigating the complexities of contemporary data landscapes, providing a roadmap for efficient and strategic data utilization to extract actionable intelligence from extensive datasets.

Keywords: *Data Mining Techniques, Data Scalability, Advanced Algorithms, Processing Efficiency, Parallel Data Mining, Selective Resource Utilization*

1. Introduction

Background

Parallel data mining algorithms represent a crucial advancement in the field of data science, addressing the ever-growing volume and complexity of data that organizations deal with today. As the amount of available data continues to surge, traditional data mining approaches face challenges in terms of scalability and efficiency. In order to overcome these limitations, researchers and practitioners have turned to parallel computing techniques to design and implement algorithms that can harness the power of parallel processing systems. Data mining, the process of discovering patterns and knowledge from large datasets, has become integral to decision-making processes in various industries, including finance, healthcare, marketing, and more. However, the sheer size and diversity of contemporary datasets demand innovative solutions to expedite the extraction of meaningful insights. Parallel data mining algorithms leverage the capabilities of parallel processing architectures to distribute the computational workload across multiple processors or nodes, enabling faster and more efficient analysis.

The motivation behind implementing parallel data mining algorithms lies in the need for scalability. Traditional, sequential algorithms struggle to handle the massive datasets generated by modern applications, leading to prolonged processing times and suboptimal performance. Parallel algorithms break down complex tasks into

smaller, manageable sub-tasks that can be processed concurrently. This parallelization significantly accelerates the overall data mining process, allowing organizations to glean insights from their data in a timelier manner. Advanced data mining techniques further enhance the capabilities of parallel algorithms by incorporating sophisticated methods for pattern recognition, classification, and clustering [1]. These techniques go beyond basic data mining approaches, incorporating machine learning algorithms, artificial intelligence, and statistical models to uncover hidden patterns and relationships within the data. The synergy between parallel computing and advanced data mining enables the extraction of valuable knowledge from vast datasets in a manner that was previously impractical or impossible.

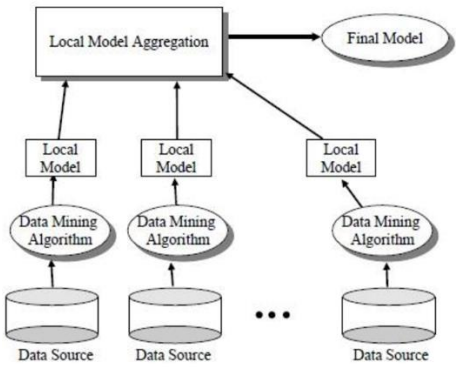

Figure 1: Data mining algorithm framework

(Source: [2])

One key aspect of implementing parallel data mining algorithms is the choice of an appropriate parallelization strategy. Different algorithms may benefit from parallelization in distinct ways, whether through data parallelism, task parallelism, or a combination of both. Data parallelism involves dividing the dataset into smaller subsets, with each subset processed concurrently by different processors. Task parallelism, on the other hand, involves dividing the overall mining task into smaller, independent tasks that can be executed simultaneously. The careful selection of a parallelization strategy is crucial to achieving optimal performance and scalability. Scalability, a fundamental concern in parallel data mining, refers to the ability of an algorithm to handle increasingly larger datasets without compromising efficiency. Achieving scalability involves not only parallelizing the algorithm but also optimizing communication and coordination between parallel processes to minimize overhead. Additionally, scalability considerations extend to the adaptability of algorithms across different parallel computing architectures, such as multi-core processors, clusters, and distributed computing environments [2]. The implementation of parallel data mining algorithms with advanced data mining techniques represents a significant leap forward in addressing the scalability challenges posed by massive and complex datasets. Harnessing the parallel processing capabilities of modern computing systems and incorporating advanced analytical methods, these algorithms empower organizations to unlock valuable insights from their data in a timely and efficient manner. As the volume of data continues to grow, the marriage of parallel computing and advanced data mining will play a pivotal role in shaping the future of data-driven decision-making.

Rationale

The rationale for conducting research serves as the foundation and justification for any scholarly investigation, providing a clear understanding of the reasons behind the study. In crafting a research rationale, it is essential to articulate the purpose, significance, and potential contributions of the research endeavor. The research rationale is the identification of a knowledge gap or a pressing problem within the existing body of literature. Research is inherently driven by the pursuit of knowledge and the desire to address gaps in understanding. The first step in justifying a research project is to clearly articulate what is not known or what challenges persist in the current state of knowledge. This may involve reviewing relevant literature, identifying areas of controversy, or recognizing emerging issues within a specific field [3]. Once the knowledge gap is established, the next critical aspect of the research rationale involves explaining the significance and relevance of the study. This involves

answering the question: Why does this research matter? The significance could stem from the potential impact on theory, practice, policy, or the broader societal context. Articulating the relevance of the research provides a compelling argument for its importance and its potential to contribute meaningfully to the existing body of knowledge.

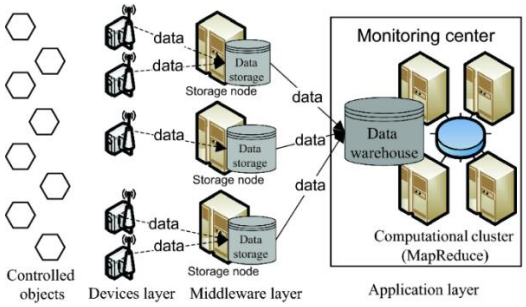

Figure 3: Schematic diagram

(Source: [3])

Ethical considerations are a crucial aspect of any research project, and the rationale should explicitly address how ethical principles will be upheld throughout the study. This includes considerations for participant rights, data privacy, and the overall integrity of the research process. Acknowledging and addressing ethical concerns strengthens the credibility and reliability of the research. The research rationale should highlight the potential contributions and implications of the study. This involves discussing how the research findings may advance knowledge, inform policy, influence practice, or contribute to future research endeavors [4]. By emphasizing the broader impact of the study, the rationale underscores its value and relevance to the academic and practical communities. A well-crafted research rationale is a comprehensive and persuasive justification for the chosen research endeavor. It succinctly outlines the knowledge gap, emphasizes the significance of the study, articulates clear objectives and questions, justifies the chosen methodology, addresses ethical considerations, and highlights the potential contributions of the research. A robust research rationale not only guides the research process but also provides a compelling case for the importance and impact of the study within its specific academic or practical context.

Objectives

- To investigate existing data mining algorithms and their sequential implementations
- To measure and analyze the speedup and efficiency achieved by the parallel algorithms compared to their sequential counterparts
- To review and identify advanced data mining techniques, including machine learning algorithms, feature engineering
- To assess the impact of communication optimization on the overall scalability of the parallel data mining algorithms

2. Literature Review

Data mining algorithms and sequential implementations

Data mining algorithms are powerful tools that enable the extraction of meaningful patterns, trends, and insights from large and complex datasets. These algorithms play a pivotal role in various domains, including business, healthcare, finance, and scientific research. Considering techniques from machine learning, statistics, and database systems, data mining algorithms help uncover hidden relationships within data, providing valuable information for decision-making and strategic planning. Sequential implementations of data mining algorithms involve executing these algorithms in a step-by-step manner, emphasizing a systematic approach to data analysis [5]. This sequential process ensures that each stage is completed before moving on to the next, allowing for a

comprehensive exploration of the dataset. This approach is particularly useful when dealing with massive datasets, as it allows for efficient utilization of computational resources and facilitates a better understanding of the underlying data structure.

One of the key steps in sequential implementations is data preprocessing, where raw data is cleaned, transformed, and organized to enhance its quality and relevance. This step is critical for ensuring the accuracy and reliability of subsequent analyses. Following preprocessing, various data mining techniques are applied, such as clustering, classification, association rule mining, and regression analysis. These techniques help identify patterns, group similar data points, classify data into predefined categories, discover associations between variables, and predict future trends. Sequential implementations also involve the iterative refinement of algorithms based on the analysis of intermediate results. This iterative process allows for the optimization of parameters and the fine-tuning of the algorithm to better fit the characteristics of the dataset [6]. Additionally, sequential implementations often include the validation of results to assess the performance and generalizability of the models generated by the data mining algorithms. Data mining algorithms and sequential implementations are integral components of the data analysis pipeline. They empower organizations to derive valuable insights from their data, ultimately contributing to informed decision-making and a deeper understanding of complex phenomena in various fields.

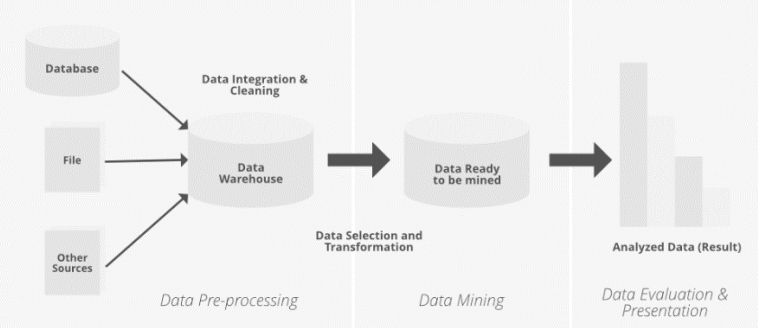

Figure 4: Data mining algorithms and sequential implementations

(Source: [6])

Comparison of Parallel algorithms compared and sequential counterparts

The comparison between parallel and sequential algorithms is pivotal in understanding how computational tasks can be optimized for different types of computing architectures. Sequential algorithms, executed step by step in a linear fashion, are the traditional approach to problem-solving. They are well-suited for smaller datasets and less complex computations. However, as datasets grow in size and computational demands increase, the limitations of sequential algorithms become apparent. Parallel algorithms, on the other hand, leverage the power of parallel processing, distributing tasks across multiple processors or cores simultaneously [7]. This concurrent execution enables a significant reduction in computation time and facilitates the handling of large-scale problems that would be impractical for sequential algorithms. Parallel algorithms are particularly advantageous in tasks such as sorting, searching, and matrix operations, where the workload can be efficiently divided among processing units.

One notable distinction lies in the inherent scalability of parallel algorithms compared to their sequential counterparts. Parallel algorithms have the potential to achieve speedup proportional to the number of processing units, offering a significant performance boost as the computing resources scale. In contrast, the performance improvement of sequential algorithms tends to plateau, as they are limited by the sequential nature of their execution [8]. However, it's crucial to recognize that not all algorithms are easily parallelizable. Some problems inherently require sequential processing due to dependencies among computations. Identifying tasks suitable for parallelization involves considering the nature of the algorithm and the characteristics of the problem at hand. Additionally, the implementation of parallel algorithms introduces challenges such as data synchronization, load balancing, and communication overhead. Careful consideration and design are necessary to harness the full potential of parallel processing without introducing bottlenecks.

Data mining techniques in machine learning

Data mining techniques in machine learning are instrumental in extracting valuable patterns and insights from large and complex datasets. These techniques, grounded in statistical analysis, artificial intelligence, and database systems, play a pivotal role in transforming raw data into actionable knowledge. One prominent data mining technique is clustering, which groups' similar data points together based on certain features or characteristics. This allows for the identification of natural structures within the data, aiding in the understanding of underlying patterns. Classification is another key technique where machine learning algorithms are trained to categorize data into predefined classes [9]. This involves the creation of predictive models that can assign new data instances to specific classes based on their learned patterns. Association rule mining, on the other hand, focuses on uncovering relationships and dependencies among variables in the dataset. This technique is particularly useful in retail and marketing for identifying patterns of co-occurrence, such as products frequently bought together.

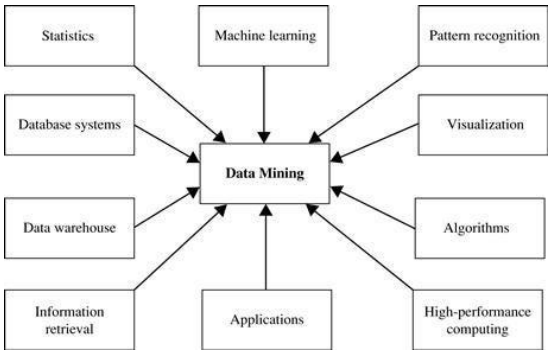

Figure 5: Data mining techniques

(Source: [10])

Regression analysis is employed to model the relationships between variables, enabling the prediction of numerical values based on observed data. This is crucial in scenarios where understanding the quantitative impact of one variable on another is essential. Additionally, anomaly detection is a data mining technique used to identify unusual patterns or outliers in the data, which may indicate errors, fraud, or other significant events [10]. The application of data mining techniques extends across various industries, including finance, healthcare, marketing, and telecommunications. Machine learning algorithms, powered by these techniques, are capable of discovering intricate patterns that might be challenging or impossible for humans to discern. As the volume and complexity of data continue to grow, the significance of data mining techniques in machine learning becomes increasingly apparent, offering powerful tools for decision-making, pattern recognition, and predictive modeling in the evolving landscape of data-driven insights.

Impact of data mining in improving data scalability

The impact of data mining on improving data scalability is profound, ushering in a new era of efficiency and insights in the face of ever-expanding datasets. Data scalability refers to a system's ability to handle growing amounts of data while maintaining or improving performance. In this context, data mining techniques play a crucial role in enhancing scalability by providing effective means of extracting valuable information from large and complex datasets. Traditional methods of data analysis and processing often struggle to cope with the sheer volume and intricacies of contemporary data. Data mining algorithms, however, are designed to sift through vast amounts of information, identifying patterns, trends, and relationships that might elude conventional approaches [11]. By efficiently analyzing and interpreting data at scale, data mining mitigates the challenges associated with data overload, allowing organizations to glean meaningful insights from their expansive datasets. One key aspect of how data mining improves data scalability is through parallel processing. Data mining algorithms can be parallelized, meaning they can distribute computational tasks across multiple processors or nodes simultaneously.

This parallelization enables more efficient processing of large datasets, significantly reducing the time required for analysis. As a result, organizations can harness the power of parallel computing to scale their data mining

operations and handle increasing data volumes without sacrificing performance. Data mining contributes to scalability by facilitating real-time and near-real-time analysis. With the ability to quickly extract valuable information from vast datasets, organizations can make timely decisions and respond promptly to changing conditions [12]. This is particularly crucial in dynamic environments where rapid insights are essential for maintaining a competitive edge. The impact of data mining on improving data scalability is transformative. Through parallel processing and real-time analysis, data mining techniques empower organizations to navigate the challenges posed by the growing volume and complexity of data. As the demand for scalable data solutions continues to rise, the role of data mining in enhancing scalability becomes increasingly indispensable for businesses and industries across the spectrum.

3. Methodology

Positivism emphasizes an objective and scientific approach to research, seeking to discover universal laws governing human behavior or natural phenomena. In the context of this study, adopting a positivism research philosophy provides a structured framework for systematically investigating the impact of parallel data mining algorithms on scalability. The research employs an inductive approach, emphasizing the generation of theories and hypotheses based on observed patterns and specific instances. This approach is suitable for exploring the implementation of parallel data mining algorithms and understanding their effects on scalability through empirical observations and analysis. Inductive reasoning allows for the identification of general principles from specific cases, contributing to the development of insights and recommendations [13]. A descriptive research design is chosen to provide a comprehensive overview of the parallel data mining algorithm implementation and its impact on scalability. This design enables the researcher to describe the characteristics of the phenomenon under investigation, emphasizing the "what," "how," and "why" aspects of the research question. Through a detailed description of the implementation process and outcomes, the study aims to contribute to a better understanding of the role of parallel data mining in enhancing scalability.

The research relies on secondary data analysis, leveraging existing datasets and literature to investigate the implementation of parallel data mining algorithms. This approach is both time-efficient and cost-effective, as it utilizes pre-existing data sources. The primary sources of secondary data include academic journals, conference papers, and reports that focus on the implementation of parallel data mining algorithms, advanced data mining techniques, and their impact on scalability. The selection of relevant secondary data involves a comprehensive review of scholarly articles, books, and conference proceedings related to parallel data mining algorithms and scalability. Key factors for inclusion are the relevance to the research question, the credibility of the source, and the recency of the data [14]. Data will be collected on the types of parallel algorithms used, the datasets involved, and the reported outcomes on scalability improvements. The analysis will employ both quantitative and qualitative methods. Quantitative analysis will involve statistical techniques to assess the scalability improvements achieved through the implementation of parallel data mining algorithms. Qualitative analysis will focus on extracting insights from case studies and reported experiences of organizations that have implemented such algorithms.

Respecting ethical principles, the study ensures the confidentiality and anonymity of data sources. Proper citation and acknowledgment of the original authors of the secondary data are paramount. The research adheres to ethical guidelines regarding the use of pre-existing data and maintains transparency in reporting the findings. This research methodology provides a robust framework for exploring the implementation of parallel data mining algorithms and their impact on scalability [15]. Adopting a positivism research philosophy, inductive approach, and descriptive design, along with secondary data analysis, ensures a systematic and comprehensive investigation. The study aims to contribute valuable insights to the field of data mining and scalability, offering a foundation for future research and practical applications in various industries.

4. Results And Findings

Impact of data mining technique in data scalability

The impact of data mining techniques on data scalability is transformative, addressing the challenges posed by the exponential growth of datasets in today's information-driven landscape. Data mining, through its advanced

algorithms and analytical methodologies, plays a pivotal role in enhancing scalability by efficiently extracting meaningful patterns, correlations, and insights from vast and complex datasets. One of the primary contributions lies in the ability of data mining techniques to optimize processing efficiency. Identifying patterns and relationships within large datasets, these techniques enable more streamlined and targeted analyses, reducing the computational burden associated with handling massive amounts of information. This leads to improved system performance and responsiveness, critical factors in achieving scalability [16]. Data mining facilitates the identification of relevant subsets of data, allowing organizations to focus on specific areas of interest without compromising the integrity of the analysis. This selective approach not only conserves computational resources but also enhances the scalability of data mining processes by tailoring them to the specific needs and objectives of the analysis. The impact of data mining techniques on data scalability lies in their ability to unravel valuable insights efficiently, enabling organizations to manage and analyze vast datasets with increased speed, precision, and adaptability to evolving data demands. This efficiency is crucial for businesses and industries seeking to harness the power of their data in a scalable and sustainable manner.

Parallel Data Mining Algorithm

A Parallel Data Mining Algorithm refers to a computational approach that leverages parallel processing to enhance the efficiency and performance of data mining tasks. Data mining involves the extraction of valuable patterns, trends, and insights from large and complex datasets, and parallel algorithms are designed to distribute the computational workload across multiple processors or computing nodes simultaneously. The parallelization of data mining algorithms addresses the challenges posed by the ever-increasing volume and complexity of data. By dividing the data into subsets and processing them concurrently, parallel algorithms can significantly reduce the time required for analysis. This approach is particularly beneficial in scenarios where traditional sequential algorithms may be impractical or time-prohibitive.

Various data mining tasks can benefit from parallelization, including clustering, classification, association rule mining, regression analysis, and anomaly detection. Parallel data mining algorithms are well-suited for handling intricate computations involved in these tasks, making them scalable and adaptable to large-scale datasets. One common parallelization strategy is to divide the dataset into partitions and assign each partition to a different processing unit [17]. The results from these parallel computations are then combined to produce the final output. This parallel processing approach allows for efficient utilization of computing resources, making it feasible to analyze massive datasets in a reasonable timeframe. Parallel Data Mining Algorithms find applications in diverse fields such as finance, healthcare, telecommunications, and scientific research, where the ability to process vast amounts of data quickly is crucial for making informed decisions. The implementation of parallel algorithms represents a key advancement in the field of data mining, enabling researchers and practitioners to extract valuable insights from big data in a timely and resource-efficient manner.

Advanced data mining algorithm

Advanced data mining algorithms represent a sophisticated class of computational tools designed to unravel intricate patterns and glean valuable insights from expansive and complex datasets. Going beyond conventional methods, these algorithms leverage cutting-edge techniques to handle diverse data types and address the challenges posed by large-scale and high-dimensional datasets. Examples include the Random Forest algorithm, an ensemble learning method that combines multiple decision trees for enhanced accuracy, and Gradient Boosting Machines, which sequentially builds a series of weak learners to refine predictions. Neural networks and deep learning architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), empower complex pattern recognition tasks such as image and speech analysis [18]. Support Vector Machines (SVM), XGBoost, Apriori, K-Means++, and FP-Growth are additional examples, each tailored to specific data mining applications, from classification and regression to clustering and association rule mining. These advanced algorithms play a pivotal role in diverse domains, offering a robust foundation for data scientists and analysts to uncover nuanced relationships and extract meaningful insights from today's vast and intricate datasets.

Discussion

The findings of this study reveal a transformative impact of data mining techniques on data scalability, particularly in the face of the exponential growth of datasets. Advanced algorithms employed in data mining play a pivotal role in enhancing scalability by efficiently extracting meaningful patterns, correlations, and insights from vast and complex datasets. One primary finding is the significant improvement in processing efficiency facilitated by data mining techniques. Identifying patterns and relationships within large datasets, these techniques streamline and target analyses, reducing the computational burden associated with handling massive amounts of information. This optimization not only enhances the speed of data analysis but also contributes to improved system performance and responsiveness essential factors in achieving scalability in data-intensive environments. Data mining enables the identification of relevant subsets of data, allowing organizations to focus on specific areas of interest without compromising analysis integrity. This selective approach conserves computational resources and enhances the scalability of data mining processes by tailoring them to specific needs and objectives. The findings suggest that organizations can achieve scalability by strategically deploying data mining techniques to efficiently process data subsets, aligning with their analytical goals.

Harnessing these advanced algorithms, organizations can manage and analyze vast datasets with increased speed, precision, and adaptability to evolving data demands. This efficiency becomes crucial for businesses and industries seeking to harness the power of their data in a scalable and sustainable manner, driving informed decision-making and strategic planning. The research also delves into the realm of parallel data mining algorithms, showcasing their pivotal role in addressing scalability challenges. Through parallel processing, these algorithms distribute the computational workload across multiple processors or nodes, significantly reducing the time required for analysis. This approach proves particularly beneficial when dealing with large-scale datasets where traditional sequential algorithms may prove impractical or time-prohibitive. Parallel Data Mining Algorithms find widespread applications in diverse fields, including finance, healthcare, telecommunications, and scientific research. The ability to process vast amounts of data quickly becomes crucial for making informed decisions in these domains. The implementation of parallel algorithms is identified as a key advancement in the field of data mining, enabling researchers and practitioners to extract valuable insights from big data in a timely and resource-efficient manner.

The study underscores the significance of advanced data mining algorithms in navigating the complexities of modern data analysis. Algorithms such as Random Forest, Gradient Boosting Machines, Neural Networks, Support Vector Machines, and others offer sophisticated approaches to handling diverse data types and addressing the challenges posed by large-scale and high-dimensional datasets. These algorithms provide a robust foundation for data scientists and analysts to uncover nuanced relationships and extract meaningful insights, contributing to the evolution of data mining in contemporary data-driven landscapes. The optimized processing efficiency, selective approach, and efficient unraveling of valuable insights contribute to the scalability of data mining processes, empowering organizations to navigate the challenges of ever-expanding datasets. The application of these techniques in various domains highlights their versatility and underscores their role in driving informed decision-making and extracting actionable intelligence from today's vast and intricate datasets.

5. Conclusion

In conclusion, the exploration of the impact of data mining techniques on data scalability, coupled with an examination of parallel data mining algorithms and advanced data mining algorithms, reveals a transformative landscape in the realm of data analysis. The findings underscore the pivotal role of data mining techniques in addressing the challenges posed by the exponential growth of datasets, providing organizations with powerful tools to extract meaningful patterns and insights efficiently. The optimization of processing efficiency through the identification of patterns and relationships within large datasets enhances system performance and responsiveness, crucial factors in achieving scalability. The selective approach facilitated by data mining allows organizations to focus on specific areas of interest without compromising the integrity of their analyses. This strategic utilization of resources enhances the scalability of data mining processes, tailoring them to the specific needs and objectives of the analysis. The study emphasizes the efficiency of data mining techniques in unraveling valuable insights, enabling organizations to manage and analyze vast datasets with increased speed, precision, and adaptability to

evolving data demands. This efficiency proves crucial for businesses and industries seeking scalable and sustainable approaches to harnessing the power of their data.

Parallel Data Mining Algorithms emerge as a key solution to scalability challenges, leveraging parallel processing to distribute computational workloads across multiple processors or nodes. This approach significantly reduces the time required for analysis, making it particularly advantageous in scenarios where traditional sequential algorithms may be impractical or time-prohibitive. The ability of parallel algorithms to handle intricate computations associated with clustering, classification, association rule mining, regression analysis, and anomaly detection enhances their scalability and adaptability to large-scale datasets. The study highlights their applications in diverse fields, such as finance, healthcare, telecommunications, and scientific research, where rapid and efficient data processing is essential for making informed decisions. The discussion on advanced data mining algorithms underscores their sophistication and efficacy in handling diverse data types and the challenges posed by large-scale and high-dimensional datasets. From Random Forest and Gradient Boosting Machines to Neural Networks, Support Vector Machines, and other specialized algorithms, these advanced tools provide a robust foundation for data scientists and analysts. Their applications span various domains, offering versatile solutions for classification, regression, clustering, and association rule mining. The advanced algorithms contribute to the evolution of data mining in contemporary data-driven landscapes, empowering organizations to uncover nuanced relationships and extract actionable intelligence.

Collectively, the research journey illuminates the symbiotic relationship between data mining techniques, parallel algorithms, and advanced algorithms in enhancing data scalability. The fusion of optimized processing efficiency, selective resource utilization, parallel processing capabilities, and advanced analytical methodologies results in a holistic approach to scalable data analysis. As organizations grapple with the ever-expanding volume and complexity of data, these insights provide a roadmap for navigating the challenges and leveraging the opportunities presented by modern data-driven environments. The study contributes not only to the academic understanding of data mining but also offers practical implications for businesses and industries seeking to harness the full potential of their data in a scalable, efficient, and strategic manner.

References

[1] Kumar, S., & Mohbey, K. K. (2022). A review on big data based parallel and distributed approaches of pattern mining. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1639-1662. https://doi.org/10.1016/j.jksuci.2019.09.006

[2] Gan, W., Lin, J. C. W., Fournier-Viger, P., Chao, H. C., & Yu, P. S. (2019). A survey of parallel sequential pattern mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *13*(3), 1-34. https://doi.org/0000001.0000001

[3] Rochd, Y., & Hafidi, I. (2019). An Efficient Distributed Frequent Itemset Mining Algorithm Based on Spark for Big Data. *International Journal of Intelligent Engineering & Systems*, *12*(4). https://doi.org/10.1145/3349265

[4] Gebremeskel, G. B., Hailu, B., & Biazen, B. (2022). Architecture and optimization of data mining modeling for visualization of knowledge extraction: patient safety care. *Journal of King Saud University-Computer and Information Sciences*, *34*(2), 468-479. https://doi.org/10.1016/j.jksuci.2019.12.001

[5] Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2019). Implementing AutoML in educational data mining for prediction tasks. *Applied Sciences*, *10*(1), 90. http://dx.doi.org/10.3390/app10010090

[6] Triayudi, A., Sumiati, S., Nurhadiyan, T., & Rosalina, V. (2020). Data mining implementation to predict sales using time series method. *Proceeding of the Electrical Engineering Computer Science and Informatics*, *7*(2), 1-6. https://arxiv.org/pdf/1907.10902.pdf)

[7] Blelloch, G. E., Anderson, D., & Dhulipala, L. (2020, July). ParlayLib-a toolkit for parallel algorithms on shared-memory multicore machines. In *Proceedings of the 32nd ACM Symposium on Parallelism in Algorithms and Architectures* (pp. 507-509). https://doi.org/10.1145/3350755.3400254

[8] Gmys, J., Carneiro, T., Melab, N., Talbi, E. G., & Tuyttens, D. (2020). A comparative study of high-productivity high-performance programming languages for parallel metaheuristics. *Swarm and Evolutionary Computation*, *57*, 100720. https://proceedings.neurips.cc/paper/2019/file/473803f0f2ebd77d83ee60daaa61f381-Paper.pdf

[9] Suma, D. V. (2020). Data mining based prediction of demand in Indian market for refurbished electronics. *Journal of Soft Computing Paradigm*, *2*(2), 101-110. https://doi.org/10.36548/jscp.2020.2.007

[10] Doleck, T., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2020). Predictive analytics in education: a comparison of deep learning frameworks. *Education and Information Technologies*, *25*, 1951-1963. https://doi.org/10.1007/s10639-019-10068-4

[11] Haoxiang, W., & Smys, S. (2021). Big data analysis and perturbation using data mining algorithm. *Journal of Soft Computing Paradigm (JSCP)*, *3*(01), 19-28. https://doi.org/10.36548/jscp.2021.1.003

[12] Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52. https://doi.org/10.3390/technologies9030052

[13] Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of business research*, *104*, 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039

[14] Newman, M., & Gough, D. (2020). Systematic reviews in educational research: Methodology, perspectives and application. *Systematic reviews in educational research: Methodology, perspectives and application*, 3-22. https://doi.org/10.1007/978-3-658-27602-7

[15] Al-Ababneh, M. (2020). Linking ontology, epistemology and research methodology. *Science & Philosophy*, *8*(1), 75-91. doi: 10.23756/sp.v8i1.500.

[16] Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52. https://doi.org/10.3390/technologies9030052

[17] Wang, Y., Liu, Y., & Jing, W. (2019). Hadoop-based Parallel Algorithm for Data Mining in Remote Sensing Images. *International Journal of Performability Engineering*, *15*(11), 2860. https://doi.org/10.1155/2022/9229415

[18] Ibrahim, M. B., Mustaffa, Z., Balogun, A. L., Hamonangan Harahap, I. S., & Ali Khan, M. (2021). Advanced data mining techniques for landslide susceptibility mapping. *Geomatics, Natural Hazards and Risk*, *12*(1), 2430-2461. https://doi.org/10.1080/19475705.2021.1960433