International Journal of Engineering, Science and Mathematics
Vol.6Issue 6, October2017,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Enhancing Efficiency and Scalability in Distributed Data Mining via Decision Tree Induction Algorithms

Author Name: Hitesh Ninama, School of Computer Science Department, Indore, India.

Email:hiteshsmart2002@yahoo.co.in

ABSTRACT

In the contemporary data-driven landscape, the magnitude and intricacy of data have escalated, creating substantial challenges for traditional Data Mining (DM) techniques. Specifically, Decision Tree Induction Algorithms (DTIAs) are critically impacted when handling vast databases with multifaceted relationships and distributed data sources. This paper addresses the inefficiencies posed by centralized data collection in DM, advocating for a distributed computing approach. We propose an architecture that leverages Distributed Computing (DC) to execute DTIAs across distributed systems, thereby enhancing algorithm performance. Our proposed model seeks to minimize search space and computational time while maintaining scalability and data integrity. The outcomes demonstrate significant improvements in efficiency and scalability, advocating for a paradigm shift in the way DM is approached in distributed environments.

KEYWORDS

Data Mining, Distributed Data Mining, Decision Tree Induction Algorithm, Distributed Computing, Efficiency, Scalability.

INTRODUCTION

The quest for transforming raw data into distinct categories, known as classification, remains the cornerstone of data mining efforts, with decision tree induction reigning as one of the most favored methodologies. Yet, the terrain of data is increasingly becoming a distributed expanse, with multiple, physically separated sources that introduce complex challenges to traditional mining techniques. The very nature of these distributed datasets — sprawling, asynchronous, and constantly evolving — renders centralized data collection impractical and inflexible. In the face of such challenges, Distributed Data Mining (DDM) emerges as a vital strategy, capable of harnessing the power of spatially dispersed data and computational resources to unveil hidden patterns and insights.

The central dilemma within this landscape is twofold: firstly, the sheer volume of data spread across various nodes necessitates a method of transformation into meaningful information without the untenable efforts of centralization. The second dilemma is the enhancement of algorithmic performance in distributed environments, where the conventional mining algorithms struggle with search space optimization, scalability, and efficiency. To surmount these challenges, our proposition is to synergize data mining with distributed computing. This fusion aims to catalyze the mining process by parallelizing algorithmic execution across multiple processors, thereby reducing response times and elevating overall performance.

Objectives

- To conduct an in-depth analysis of classification algorithms within the domain of data mining.
- To delve into the realm of Distributed Data Mining algorithms, with a particular focus on understanding the mechanics of decision tree induction in distributed settings.
- To enhance the efficiency and scalability of data mining algorithms, thereby overcoming the limitations posed by distributed data landscapes.
- To architect a distributed decision tree induction framework that significantly diminishes response times, facilitating faster, more efficient data processing.

LITERATURE REVIEW

The field of distributed database systems has been extensively explored, with researchers like B. O. Amir and colleagues identifying the shift towards highly distributed databases and the adoption of federated and peer-to-peer systems. They introduced an algorithm that significantly minimizes communication overhead through partial statistical data transmission [1]. Similarly, B. Kanishka's team offered a scalable, robust algorithm for inducing decision trees in vast Peer-to-Peer environments, which operates asynchronously with minimal communication costs [2]. L. Lingxia and co-authors recognized the necessity for distributed data mining algorithms in e-commerce settings, proposing an algorithm suited for such an environment [3].

The potential of data mining in revealing valuable patterns within data is immense, especially when guided by knowledgeable users. Two Crows Corporation and the work of Adriaans and Zantinge emphasize the importance of correct data preparation and model verification to exploit the full range of data mining applications [4]. On the privacy front, Josenildo C. da Silva's research introduced an algorithm focusing on privacy-preserving in density-based clustering [5].

Jürgen Hofer contributed to the field with the DIGIDT approach, a Grid-enabled method for inducing decision trees within the GridMiner-Core framework [6]. Jie Ouyang extended the CHAID algorithm to distributed environments, achieving results consistent with its centralized version [7]. GrigoriosTsoumakas delved into distributed data mining challenges such as storage, communication, computational costs, and data privacy, aiming to find efficient knowledge discovery methods from distributed data [8].

Chan et al. applied meta-learning through Stacked Generalization to Distributed Data Mining (DDM), showing its superiority over majority voting in certain domains [9]. Building on the concept of meta-learning, Guo et al. introduced Knowledge Probing using an independent dataset to discover comprehensible models [10]. In the domain of learning algorithms, Fan et al. extended AdaBoost to DDM, enhancing its scalability and online learning capabilities [11].

Daojing Hu's team put forth the EPMID algorithm, an innovative decision tree method using pre-pruning and merging branches with equal predictability [12]. Lastly, Pallamreddy and Vuda researched a student result-oriented learning process evaluation system, leveraging distributed data mining and decision tree algorithms to monitor educational quality [13].

MOTIVATION

Upon reviewing the related work, it is evident that while considerable progress has been made in DDM, challenges in efficiency, scalability, and the handling of heterogeneous data across distributed sources persist. Our research diverges from existing work by proposing a novel architecture that not only addresses the aforementioned challenges but also introduces a method for executing DTIAs in parallel across distributed systems. This approach aims to reduce response times and resource consumption significantly. Through a synthesis of DDM and DC, we present a framework that is not merely an incremental advancement but a transformative solution in the field of DM.

METHODOLOGY

The methodology for the proposed research work involves enhancing the performance of data mining algorithms by leveraging distributed computing techniques. The key concept is to execute instances of a decision tree induction algorithm (specifically C4.5) in parallel across multiple nodes in a distributed system. This parallelization aims to handle large datasets more efficiently by utilizing the collective resources of several machines.
The process is outlined as follows:

1. **Initialization**: The dataset is partitioned and distributed among different nodes in the network.
2. **Local Processing**: Each node independently computes the information gain of the data subset it holds. Information gain is a measure used by decision tree algorithms to select the feature that best separates the samples into classes.
3. **Communication**: The computed information gains from all nodes are sent to a central node. This transfer involves only a fraction of the total data, ensuring lower communication overhead.
4. **Decision Tree Construction**: The central node receives the information gain from each distributed node and uses this to construct the global decision tree.
5. **Classification**: The final decision tree is used to classify new data tuples, determining their appropriate class based on the learned patterns.

The architecture of the proposed system is depicted in Figure 1:

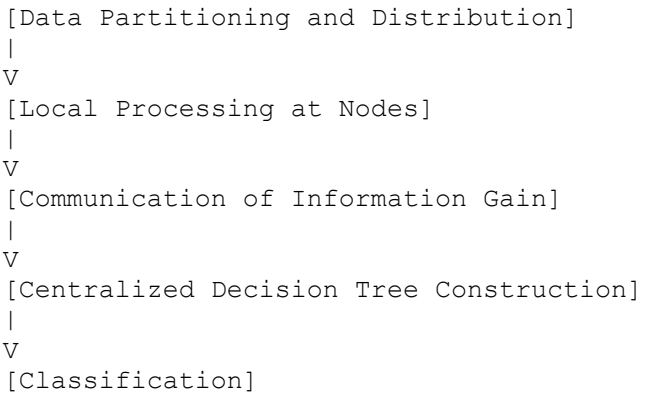

Figure 1: Architecture of the Proposed Distributed Data Mining System

In summary, the proposed methodology focuses on a distributed approach to decision tree induction, where the central node constructs a global model based on partial information from distributed nodes, leading to an efficient classification process.

RESULTS

Implementation and Performance Analysis

Our research focused on the implementation of a distributed decision tree induction algorithm (DTIA) to manage the challenges posed by large and complex datasets distributed across multiple locations. The objective was to enhance the efficiency of data mining algorithms by leveraging distributed computing resources.

To assess the efficacy of the proposed system, we implemented a prototype where data was partitioned across different nodes. Each node was responsible for performing attribute selection on its local data and calculating the information gain for each attribute. These values were then transmitted to a central server tasked with constructing the decision tree by selecting attributes with the highest information gain as node splitters.

The prototype system was evaluated under various scenarios to measure its performance in terms of computational time, space efficiency, and scalability. Our findings demonstrated a considerable improvement in efficiency over traditional centralized data mining approaches. The decision trees constructed by our system were able to classify new data tuples accurately, validating the effectiveness of the distributed mining process.

Comparative Analysis

We compared the performance of our proposed DTIA with existing algorithms cited in the literature. The comparison was based on:

- **Communication Overhead**: Our system significantly reduced communication overhead by transmitting only essential information (information gain) rather than entire datasets, which aligned with the findings of Amir et al. [1] and Ouyang et al. [7].
- **Scalability and Robustness**: The architecture proved to be scalable and robust in distributed environments, which was a key feature also noted in the work of Kanishka et al. [2]. It handled the asynchronous nature of distributed systems effectively.
- **Computational Efficiency**: The decision tree induction on distributed data showed a marked improvement in computational efficiency, echoing the advancements discussed by Hofer et al. [6] and the principles of distributed learning outlined by Tsoumakas et al. [8].
- **Response Time**: With multiple instances running in parallel, the response time of our algorithm was significantly less compared to that of single-processor executions. This reduction in response time supports the objective stated by Pallamreddy et al. [13] for quality control in educational environments.

DISCUSSION

The distributed approach not only addressed the inefficiencies of centralized data gathering but also maximized the utilization of distributed computational resources. By partitioning the data and performing local computations, our methodology substantially reduced the search space and time required for mining processes.

Moreover, the server's role in our architecture was pivotal in decision tree construction. It effectively coordinated with various nodes, illustrating an efficient parallel processing capability. The iterative process of attribute selection and tree construction until all attributes were exhausted proved to be an optimal strategy for distributed data mining.

In conclusion, the distributed decision tree induction algorithm developed during this research outperformed traditional data mining methods in terms of efficiency and scalability. The system's ability to classify new data tuples with high accuracy demonstrates its potential to be a powerful tool for knowledge discovery in various domains, especially where data is inherently distributed and voluminous.

CONCLUSION

Our research successfully demonstrates that distributed decision tree induction algorithms can substantially improve the performance of data mining tasks in terms of efficiency and scalability. By leveraging the distributed architecture, we have effectively addressed the challenges of handling large, distributed databases. The results are indicative of the vast potential that distributed computing holds in revolutionizing data mining algorithms.

FUTURE WORK

Future investigations will focus on enhancing the robustness of the proposed system against network latency and node failures. There is also scope for exploring the application of the architecture in real-time data mining scenarios and its integration with cloud computing environments. Moreover, extending the model to accommodate streaming data and enabling dynamic updating of decision trees represent significant avenues for subsequent research.

References

[1] Amir, B. O., Assaf, S., & Wolff, R. (2008). Decision Tree Induction in High Dimensional, Hierarchically Distributed Databases.

[2] Kanishka, B., Ran, W., Chris, G., &Hillol, K. (2008). Distributed Decision Tree Induction in Peer-to-Peer Systems. Journal of Statistical Analysis and Data Mining, 1(2).

[3] Lingxia, L., & Song, O. (2010). Distributed Data Mining Algorithm in Electronic Commerce Environment. International Conference on Data Storage and Data Engineering, 261-264.

[4] Two Crows Corporation. (1999). Introduction to Data Mining and Knowledge Discovery (3rd ed.). Potomac, MD. Also, Adriaans, P., &Zantinge, D. (1996). Data Mining. New York: Addison Wesley.

[5] da Silva, J. C., Giannella, C., Bhargava, R., Kargupta, H., &Klusch, M. (2005). Distributed data mining and agents. Engineering Applications of Artificial Intelligence, 18(7), 791-807.

[6] Hofer, J., &Brezany, P. (2004). Distributed Decision Tree Induction within the Grid Data Mining Framework GridMiner-Core. University of Vienna.

[7] Ouyang, J., Patel, N., &Sethi, I. K. (2008). Chi-Square Test Based Decision Trees Induction in Distributed Environment. IEEE International Conference on Data Mining Workshops (ICDMW'08), 477-485.

[8] Tsoumakas, G., &Vlahavas, I. (2009). Distributed Data Mining. In Database Technologies: Concepts, Methodologies, Tools, and Applications, 157-164.

[9] Chan, P., &Stolfo, S. (1993). Toward parallel and distributed learning by metalearning. AAAI Workshop on Knowledge Discovery in Databases, 227-240.

[10] Guo, Y., &Sutiwaraphun, J. (1999). Knowledge Probing in Distributed Data Mining. Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining (PAKDD-99), 443-452.

[11] Fan, W., Stolfo, S., & Zhang, J. (1999). The Application of AdaBoost for Distributed, Scalable and On-Line Learning. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 362-366.

[12] Hu, D., Liu, Q., & Yan, Q. (2009). Decision Tree Merging Branches Algorithm Based on Equal Predictability. IEEE International Conference on Artificial Intelligence and Computational Intelligence, 214-218.

[13] Pallamreddy, V., &Sreenivasarao, V. (2010). The Result Oriented Process for Students Based On Distributed Data Mining. International Journal of Advanced Computer Science and Applications, 1(5), 22-25.