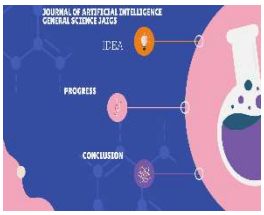
Vol.4, Issue 1, may, 2024
Journal of Artificial Intelligence General Science JAIGS

https://ojs.boulibrary.com/index.php/JAIGS



Scalable Machine Learning Solutions for Heterogeneous Data in Distributed Data Platform

Chandrashekar Althati[1], Manish Tomar[2], Jesu Narkarunai Arasu Malaiyappan[3]

[1]Medalogix, USA

[2]Citibank, USA

[3]Meta Platforms Inc, USA

ARTICLEINFO
Article History:
Received:01.05.2024
Accepted:
15.05.2024
Online: 30.05.2024
Keyword: Scalable Machine Learning, Heterogeneous Data, Distributed Data Platform, Data Variety, Distributed Processing

ABSTRACT

As the volume and variety of data continue to expand, the need for scalable machine learning solutions becomes increasingly vital, especially in distributed data platforms handling heterogeneous data sources. This research explores methods and techniques for developing scalable machine learning solutions tailored to the challenges posed by heterogeneous data in distributed environments. By addressing issues such as data variety, scalability, and distributed processing, this study aims to provide insights into building robust machine learning models capable of handling diverse data types efficiently. Through experimentation and analysis, the research seeks to uncover effective strategies for implementing scalable machine learning solutions in distributed data platforms, thereby facilitating improved data processing and decision-making capabilities across various domains.

Introduction

In today's era of big data, the proliferation of diverse data types poses a significant challenge for machine learning (ML) algorithms, particularly within distributed data platforms. The need to process heterogeneous data efficiently while maintaining scalability and performance has become increasingly vital for various industries. This paper focuses on addressing these challenges by developing scalable ML algorithms, optimizing models for enhanced performance, and exploring distributed processing methods for heterogeneous data within distributed data platforms.

The rapid expansion of data sources, including structured, semi-structured, and unstructured data, has necessitated the development of ML algorithms capable of handling diverse data types. Traditional ML approaches often struggle to efficiently process heterogeneous data within distributed data platforms due to scalability issues and performance bottlenecks. Therefore, there is a pressing need to develop scalable solutions that can effectively handle the complexities of diverse data types while operating within distributed computing environments.

To address these challenges, this research aims to develop scalable ML algorithms tailored to the requirements of distributed data platforms. These algorithms will leverage distributed computing frameworks to ensure efficient processing of heterogeneous data types, including text, images, sensor data, and more. By harnessing the power of distributed computing, these algorithms will be capable of handling large volumes of data and delivering timely insights for decision-making processes.

Furthermore, this research will investigate techniques for optimizing ML models to enhance their performance and adaptability within distributed computing environments. By fine-tuning model parameters, leveraging parallel processing capabilities, and implementing advanced optimization techniques, the aim is to improve the overall efficiency and effectiveness of ML solutions operating within distributed data platforms.

Additionally, this paper will explore methods for distributed processing of heterogeneous data, considering factors such as data partitioning, communication overhead, and fault tolerance. By examining different distributed processing strategies and their impact on scalability and efficiency, this research seeks to identify best practices for designing ML solutions within distributed data platforms.

Overall, the objectives of this research are to develop scalable ML algorithms, optimize model performance, and explore distributed processing methods for heterogeneous data. By achieving these objectives, this study aims to advance the capabilities of ML within distributed data platforms and facilitate more efficient and effective data analysis for various applications and industries.

Objectives:

1. Develop scalable machine learning algorithms capable of efficiently processing heterogeneous data types within distributed data platforms.

2. Investigate techniques for optimizing machine learning models to enhance performance and adaptability in distributed computing environments.

3. Explore methods for distributed processing of heterogeneous data to improve the scalability and efficiency of machine learning solutions within distributed data platforms.

Research Methodology:

1. Data Collection: The research will involve gathering datasets that represent heterogeneous data types commonly encountered in real-world scenarios. This may include structured data from databases, unstructured data from text documents, multimedia data such as images and videos, and streaming data from sensors or IoT devices. Datasets will be selected to cover a diverse range of data types and characteristics.

2. Algorithm Development: Scalable machine learning algorithms will be developed or adapted to efficiently process heterogeneous data within distributed data platforms. This will involve designing algorithms that can handle various data types, accommodate distributed computing architectures, and optimize performance for large-scale datasets. Common ML techniques such as supervised learning, unsupervised learning, and deep learning may be employed.

3. Model Optimization: Techniques for optimizing machine learning models in distributed computing environments will be investigated. This may include parameter tuning, feature selection, model compression, and parallelization strategies to improve performance, reduce computational overhead, and enhance adaptability to distributed data platforms.

4. Experimental Design: A series of experiments will be conducted to evaluate the performance of the developed algorithms and optimization techniques. Experiments will be designed to measure factors such as accuracy, scalability, processing time, resource utilization, and fault tolerance under various conditions. Benchmark datasets and simulation environments will be used to ensure reproducibility and rigor in the experimental process.

5. Performance Evaluation: The performance of scalable machine learning solutions for heterogeneous data in distributed data platforms will be evaluated based on predefined metrics and criteria. Comparative analysis will be

conducted to assess the effectiveness of the proposed algorithms and optimization techniques compared to existing approaches. Results will be analyzed to identify strengths, limitations, and areas for improvement.

6. Validation and Validation: The developed algorithms and optimization techniques will be validated through real-world use cases or simulations that mimic real-world scenarios. Validation will involve testing the solutions in practical settings to assess their applicability, robustness, and scalability in handling diverse data types within distributed data platforms.

7. Documentation and Reporting: Throughout the research process, detailed documentation of methodologies, experimental procedures, and results will be maintained. The findings will be reported in research papers, conference presentations, and potentially in open-access repositories to contribute to the broader research community.

Literature Review

Scalable machine learning solutions for heterogeneous data in distributed data platforms can benefit from advancements in federated learning. Challenges like data heterogeneity, communication load, and mobility can be addressed through innovative approaches like hierarchical federated learning (HFL) [1], adaptive mixing aggregation (AMA) for heterogeneous data distribution, staleness-based weighting for dynamic wireless conditions, and CPU-friendly computation-reduction schemes [2]. Additionally, FedMix, a new algorithm adjusting existing federated learning methods, shows promise in improving performance by focusing on client-side adjustments [3]. Furthermore, strategies like serial pipeline training (SPT) and global knowledge regularization (GKR) in FedSPARK enhance the efficiency of federated learning by reducing client computation and communication while improving model performance on heterogeneous data [4] [5]. These approaches collectively contribute to scalable and effective machine learning solutions for handling heterogeneous data in distributed platforms.

Background

Intelligent automotive technology is rapidly advancing, with autonomous car navigation on the brink of becoming a reality. One emerging trend in data security is blockchain technology, which holds promise for various applications. We foresee a future where all vehicles will feature fully equipped onboard computers capable of installing secure applications, including navigation and sensor reading functionalities, without the need for additional hardware modifications.

The rise of cloud computing and edge computing has led to an exponential growth in data, particularly within vehicular social networks. This data influx presents opportunities to enhance the security, convenience, and entertainment features of applications within these networks. Effective data analysis methods, such as machine learning and deep learning, play a crucial role in harnessing this data deluge.

Among machine learning techniques, the support vector machine (SVM) model stands out for its performance and robustness, making it widely applicable across various domains. For instance, in vehicular social networks, numerous entities, including vehicle manufacturers, management agencies, and application service providers, possess diverse datasets that complement each other in terms of attributes. This scenario gives rise to heterogeneous data.

However, individual organizations' datasets often lack multidimensional coverage, limiting their utility. Particularly in SVM classifier training, the quality of the dataset significantly impacts the classification outcome. Hence, sharing heterogeneous data among multiple institutions becomes imperative. Through data sharing, datasets with diverse attributes can be merged to enhance classifier effectiveness.

On another note, the fused heterogeneous data can be vertically partitioned into sub-datasets based on attributes provided by each unit. Yet, data sharing poses serious privacy challenges. Heterogeneous data often contains users' sensitive information, leading to increasing regulatory constraints on data sharing. Additionally, data owners highly value the privacy of their data, making direct sharing less feasible.

Privacy disclosure has long been a prominent issue in various scenarios, with considerable attention directed towards securely training machine learning classifiers over horizontally and vertically partitioned datasets. Many existing solutions utilize secure multi-party computation (SMC) to mitigate privacy risks. However, these schemes often struggle to strike a balance between security and efficiency. Additionally, relying on one or more trusted servers during the training process is impractical in real-world scenarios.

To address these challenges, we propose an efficient and secure SVM classifier training scheme based on consortium blockchain, eliminating the need for third-party involvement. Our approach tackles the issues associated with privacy protection and real-world applicability.

In our proposed mechanism, we leverage homomorphic encryption instead of the traditional differential privacy protection scheme, as it ensures secure training without compromising data privacy. By integrating blockchain technology, we establish a decentralized data sharing platform where participants can securely share their data. The platform's access control and permission mechanisms ensure the confidentiality of external data and the transparency of internal data.

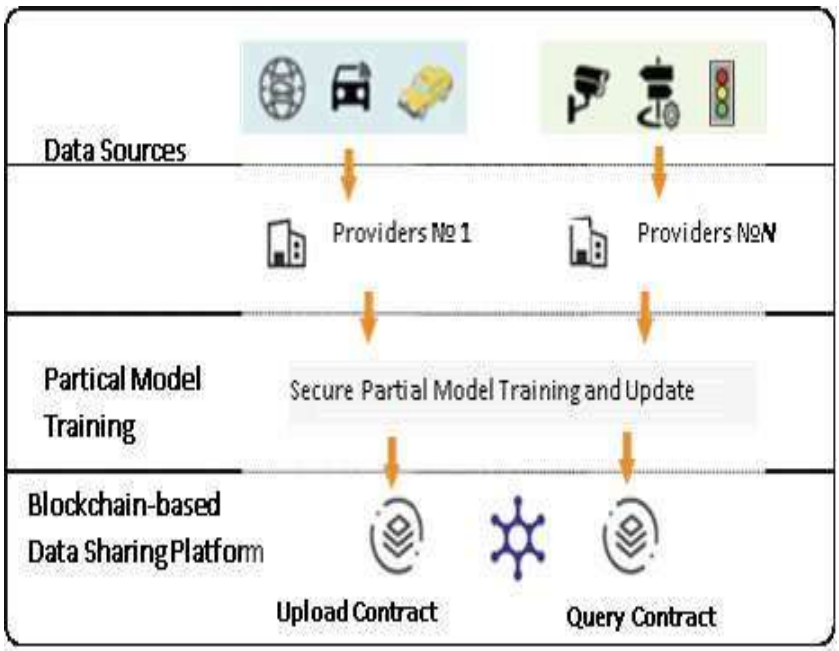
Our SVM training scheme enhances the security and efficiency of heterogeneous data sharing. We establish an open, reliable, and transparent data sharing platform using blockchain, eliminating the need for trusted third parties. Participants perform most of the training locally based on plaintext data, enhancing privacy protection. We introduce a threshold homomorphic encryption scheme to protect shared data in a decentralized environment while preserving its homomorphic property. The scheme allows for customizable privacy protection levels by adjusting the threshold size. Extensive experiments conducted on real datasets validate the feasibility and efficiency of our approach.

Secure Machine Learning over Heterogeneous Data

Let's consider a dataset, \(D \), composed of several participants, each having its own dataset denoted as \(D_{pp} \) where \(p \) ranges from \(A \) to \(N \). In this context, \(x_p \) represents the \(i \)-th instance in \(D_p \), while \(y_i \) is shared as a data label among all related instances \(x_i \). When training a SVM classifier, we define \(w \) as the model parameters, \(\Delta t \) as the gradient in the \(t \)-th iteration, and \(\lambda \) as the learning rate. Additionally, we assume \([[m]] \) to represent the encryption of message \(m \) under Paillier. Table 1 outlines the notations used in this paper.

Table1.Notations

Notations	Description
D^A	The data set of participant
d^A	The dimension of data set D^A
x_i^A	The data instance of data set D
y_i	Number



System Model

Our system is categorized into three components based on their interaction with the data, as illustrated in Fig. 1: the Data Device (DD), Data Provider (DP), and Blockchain Service Platform (BSP).

- Data Device: This refers to devices capable of data generation, such as sensors, mobile devices, and others. These devices collect valuable data, which is then processed for further analysis.

- Data Provider: Data providers are entities responsible for generating, collecting, and storing data from different sources. These participants, or data providers, contribute varied data sets due to differences in equipment and data processing methods. These diverse data sets complement each other in terms of attributes. Besides serving as data providers, these participants also engage as model trainers in collaborative machine learning efforts. In this paper's scheme, most of the training tasks are performed locally by the participants.

- Blockchain Service Platform: This platform operates on the consortium blockchain. It offers a transparent data sharing platform distributed among the participants, enabling them to access all data recorded within the BSP. Additionally, the BSP ensures the integrity of data records, preventing unauthorized alterations. Moreover, it features robust security measures, rendering data outside the participants' domain invisible. Communication between the BSP and participants is encrypted, ensuring data confidentiality and preventing leakage.

Threat Model

In our scheme, there is a singular role of the data provider. We consider participants to be honest but curious within the security model, implying that while all participants are curious about the data of others, they will abide by the rules of the scheme. Moreover, given the extensive interactions between participants and the BSP, potential threats during this interaction process are also taken into account.

- Known Ciphertext Model: The BSP serves as a common and transparent data sharing platform for all participants. Data shared by each participant is visible to others, including dense intermediate values and decrypted calculation results.

- Known Background Model: Under this model, we assume that multiple participants may conspire and collaborate to analyze shared data, enabling them to obtain more information compared to the previous threat model.

Under these system and threat models, we establish three system design goals to meet the requirements of security, accuracy, and performance:

- Full Data Privacy Protection: Throughout the entire training process, under both threat models, the privacy of the original data and shared intermediate values remains intact, preventing participants from inferring valuable information. Additionally, data within the sharing platform is guaranteed to be invisible to external entities.

- High Accuracy of Training Results: While privacy protection schemes may introduce noise into the calculation process, our design goal is to obtain a classifier that is comparable in accuracy to conventional training conditions.

- Low Training Overhead: Privacy protection schemes may increase training overhead, primarily due to additional computing operations like encryption and decryption, as well as additional communication overhead. Therefore, our solution aims to ensure low training overhead while maintaining security.

Secure SVM Training Scheme over Heterogeneous Datasets

In this section, we outline a secure SVM training scheme involving three participants to elucidate the role of each participant in the training process. Assuming each participant contributes a training set with complementary attributes, the training process comprises three primary stages: local training, gradient update judgment, and model update, as depicted in Fig. 3. These stages entail two data sharing operations and one decryption operation. Through multiple iterations, each participant obtains a partial model, subsequently uploaded to the blockchain to collectively form a complete model.

Data privacy protection in this solution relies on a threshold homomorphic encryption algorithm. Prior to model training, each participant generates a pair of public and private keys, where the public key is uniform, and the private key varies. Utilizing a secret sharing scheme combined with an existing threshold key management system, these key pairs are negotiated and distributed. Furthermore, all three participants join the consortium blockchain data sharing platform as nodes, necessitating identity authentication before participation. Lastly, participants initialize model parameters and preprocess datasets, including standardized labeling and sample ordering.

Data Sharing on BSP and Security Analysis

Participants rely on BSPs to securely compute intermediate values, simplifying complex point-to-point communication. Data is managed on-chain, and queries are executed via smart contracts. Throughout the iteration process, each participant uploads data twice: once for calculating the intermediate value (IV) and once for the decrypted value (DV), with corresponding reads.

1. The Format of IVs

- Iteration Round: Indicates the round of data exchange in collaborative model training, managed by smart contracts.

- DP ID: Identifies the data owner, automatically recorded upon data upload.

- Training Intermediate Value: Encrypted state's intermediate value during model training. Participant-provided values are summed and compared to 1 in the encrypted state.

- r1: Unencrypted positive integer used for comparison.

- r2: Encrypted positive integer used for comparison.

- r3: Unencrypted positive integer used for comparison.

- Random Positive Integer: Generated randomly by each participant, determining data instances selected in the next iteration.

2. The Format of DVs

- Iteration Round: Similar to IVs, indicating the round of data exchange.

- DP ID: Identifies the data owner.

- Decrypted Value: Each participant decrypts the result based on their private key, allowing them to collectively obtain the final decryption result.

Conclusion

This section presents a robust and secure SVM training scheme tailored for multiple data providers engaged in training SVM classifiers on vertically partitioned datasets. Our approach integrates consortium blockchain technology and threshold Paillier encryption to establish a decentralized and secure SVM training platform. With a focus on high performance, the majority of training operations are conducted locally on raw data, minimizing the need for sharing intermediate values across platforms.

References List:

[1]. Soni, J., Gangwani, P., Sirigineedi, S., Joshi, S., Prabakar, N., Upadhyay, H., & Kulkarni, S. A. (2023). Deep Learning Approach for Detection of Fraudulent Credit Card Transactions. In Artificial Intelligence in Cyber Security: Theories and Applications (pp. 125-138). Cham: Springer International Publishing.

[2]. Verma, V., Vutukuru, K. S., Divvela, S. S., & Sirigineedi, S. S. (2022). Internet of things and machine learning application for a remotely operated wetland siphon system during hurricanes. In Water Resources Management and Sustainability (pp. 443-462). Singapore: Springer Nature Singapore.

[3]. Talati, D. (2023). Telemedicine and AI in Remote Patient Monitoring. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 254-255.

[4]. Talati, D. (2024). Virtual Health Assistance–AI-Based. Authorea Preprints.

[5]. Talati, D. (2023). Artificial Intelligence (Ai) In Mental Health Diagnosis and Treatment. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 251-253.

[6]. Talati, D. (2024). Ethics of AI (Artificial Intelligence). Authorea Preprints.

[7]. Talati, D. V. AI Integration with Electronic Health Records (EHR): A Synergistic Approach to Healthcare Informatics December, 2023.

[8]. Singla, A., & Malhotra, T. (2024). Challenges And Opportunities in Scaling AI/ML Pipelines. Journal of Science & Technology, 5(1), 1-21.

[9]. Singla, A., & Chavalmane, S. (2023). Automating Model Deployment: From Training to Production. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 340-347.

[10]. Gehrmann, S., & Rončević, I. (2015). Monolingualisation of research and science as a hegemonial project: European perspectives and Anglophone realities. Filologija, (65), 13-44.

[11]. Roncevic, I. (2021). Eye-tracking in second language reading. Eye, 15(5).

[12]. Šola, H. M., Gajdoš Kljusurić, J., & Rončević, I. (2022). The impact of bio-label on the decision-making behavior. Frontiers in sustainable food systems, 6, 1002521.

[13]. Sirigineedi, S. S., Soni, J., & Upadhyay, H. (2020, March). Learning-based models to detect runtime phishing activities using URLs. In Proceedings of the 2020 4th international conference on compute and data analysis (pp. 102-106).

[14]. Verma, V., Bian, L., Ozecik, D., Sirigineedi, S. S., & Leon, A. (2021). Internet-enabled remotely controlled architecture to release water from storage units. In World Environmental and Water Resources Congress 2021 (pp. 586-592).

[15]. Soni, J., Sirigineedi, S., Vutukuru, K. S., Sirigineedi, S. C., Prabakar, N., & Upadhyay, H. (2023). Learning-Based Model for Phishing Attack Detection. In Artificial Intelligence in Cyber Security: Theories and Applications (pp. 113-124). Cham: Springer International Publishing.

[16]. Talati, D. (2023). AI in healthcare domain. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 256-262.

[17]. Talati, D. (2024). AI (Artificial Intelligence) in Daily Life. Authorea Preprints.