Big data analysis methods based on artificial intelligence technology

202422900232 王冉恒

Abstract： With the rapid development of information technology, big data has become the core driving force for social progress and industrial transformation. The popularization of industrial Internet of Things, cloud computing, and sensor networks has led to an exponential growth in data volume. Traditional data processing methods are struggling to cope with the challenges of massive, high-dimensional, and heterogeneous big data. Against this backdrop, artificial intelligence (AI) technology, with its powerful learning and reasoning capabilities, has become the core tool for big data analysis. AI technologies such as machine learning (ML), deep learning (DL), and computational intelligence (CI), supported by distributed computing frameworks like Hadoop and Spark, have made significant progress in areas such as data mining, pattern recognition, and predictive analysis. This article reviews the latest advancements in big data analysis methods based on AI technology and discusses existing challenges and future research directions.[1]

I. Big Data Analysis Methods Based on Machine Learning

Machine learning, as the most core branch of artificial intelligence, is widely applied in various tasks of data mining[2]. The following elaborates on the application of machine learning methods in the big data environment from four main directions: big data clustering, association analysis, classification, and prediction：

1. Big Data Clustering

Clustering analysis is an important means to discover the inherent distribution patterns of samples from big data. Traditional clustering algorithms such as K-means, hierarchical clustering, and density-based clustering methods face problems such as high computational complexity and large storage requirements when dealing with massive data. To address these challenges, researchers have parallelized these algorithms through distributed computing platforms[3]. For instance, implementing the K-means clustering algorithm using the MapReduce framework not only enables the division of the data set into multiple data blocks for parallel processing but also improves clustering efficiency and accuracy by reducing the number of iterations and optimizing the selection of initial centers[4].

In addition, for clustering problems involving unstructured and heterogeneous data, improved methods based on density clustering (such as DBSCAN) have also been applied in the big data environment. Some studies have achieved clustering analysis of trajectory data by introducing fast search strategies and dynamic time warping algorithms, effectively enhancing the robustness and accuracy of complex data processing[5].

2. Big Data Association Analysis

Association rule mining mainly studies the intrinsic relationships between data items. Traditional association analysis algorithms include the Apriori algorithm and the FP-Growth algorithm. Due to the need for multiple database scans in mining frequent itemsets in massive data, single-machine serial processing is inefficient[6]. Therefore, the parallelization of algorithms in a distributed environment has become an inevitable trend. Parallel Apriori and FP-Growth algorithms based on MapReduce and Spark platforms, through data partitioning, local frequent pattern tree construction, and load balancing strategies, have effectively reduced I/O load and computing latency, improving mining efficiency. Furthermore, for the mining of rare rules in big data, studies have adopted new methods such as the Apriori inverse algorithm and negative rough association rule algorithm, further expanding the application scenarios of big data association rules, such as logistics transportation, network security, and fault diagnosis[7].

3. Big Data Classification

Big data classification mainly involves the automatic classification of data from different fields, with application scenarios including network intrusion detection and medical diagnosis. Researchers have implemented traditional classification methods such as decision trees, random forests, and support vector machines in parallel using the MapReduce distributed framework, solving the problem of tight computing resources in the big data environment[8]. For example, the MapReduce-based random forest algorithm and the decision tree model based on cost-sensitive learning strategies have not only improved classification accuracy but also reduced training time. Meanwhile, with the continuous enrichment of data types, researchers are constantly exploring and improving algorithms for classification problems involving imbalanced data and high-dimensional features, such as combining the K-nearest neighbor algorithm with distributed computing to build efficient and robust classifiers[9].

4. Big Data Prediction

Big data prediction is another key task in big data analysis and is widely applied in fields such as finance, energy, healthcare, and manufacturing. Prediction methods based on machine learning, such as logistic regression, regression trees, and support vector machines, can capture trends and change patterns in large-scale data. Studies have shown that by conducting correlation analysis on multiple data sources, scalable transaction models, energy load prediction models, and disease prediction models can be established, thereby providing data support for real-time decision-making.

II. Big Data Analysis Methods Based on Deep Learning

Since Geoffrey Hinton and others proposed the theory of deep learning, it has become a popular research direction in the field of artificial intelligence. Deep learning models have performed well in areas such as image processing, speech recognition, and natural language processing, and their application in big data analysis also holds great potential.

1. Distributed Implementation of Deep Learning Based on MapReduce

The training process of deep learning models is usually computationally intensive and requires numerous iterations. To reduce training time costs, researchers have utilized the MapReduce framework to parallelize common models such as deep belief networks, restricted Boltzmann machines, and BP neural networks. For instance, by employing batch updates and data parallel strategies, distributed training of weights in each layer of deep networks has been achieved, significantly reducing training time. Additionally, by improving the mapping mechanism, the shortcomings of traditional MapReduce in iterative computations have been addressed, enhancing overall convergence speed[10].

2. Distributed Implementation of Deep Learning Based on Spark

Compared to MapReduce, the Spark platform, with its focus on in-memory computing, is more suitable for iterative tasks. To address the high number of iterations in deep learning training, researchers have implemented parallel training of convolutional neural networks (CNNs), deep Boltzmann machines, and other deep models on the Spark platform[11]. Through Spark's RDD partitioning mechanism, not only has the data loading and preprocessing process been accelerated, but techniques such as batch normalization and multi-cross-validation have also been utilized to further optimize model performance. Some application cases have demonstrated that this Spark-based deep learning approach exhibits high accuracy and stability in tasks such as image recognition and mobile data behavior analysis[12].

Ⅲ. Current Status and Challenges of AI Technology in Big Data Analysis

Big data analysis methods based on machine learning and deep learning have been applied in multiple fields including industrial manufacturing, intelligent transportation, financial risk control, and medical diagnosis. In the field of intelligent manufacturing, through the analysis of data throughout the product life cycle, enterprises can achieve product quality monitoring, process optimization, and supply chain management; in the financial sector, real-time data prediction and correlation analysis have made risk warnings and investment decisions more precise[13]. However, despite numerous breakthroughs, current research still faces the following main challenges：

Data preprocessing and sample quality: Big data often comes from diverse sources and has complex formats. How to remove redundancies and improve data quality during the preprocessing stage remains a prerequisite for big data analysis. Specialized preprocessing strategies are needed for the characteristics and noise interference issues of data in different fields[14].

Algorithm parallelization and resource scheduling: Although distributed platforms can accelerate model training, how to rationally partition data blocks, balance computing loads, and reduce communication overheads remains an important factor affecting parallel efficiency. Particularly in deep learning, the synchronization issue of parameter updates in each layer needs further optimization.

Model Generalization and Interpretability: Currently, most big data analysis models focus on improving prediction accuracy. However, how to enhance the generalization ability of models in different data environments and provide reasonable explanations for the results remains a challenging issue. Especially in high-risk fields such as finance and healthcare, interpretability has become an important consideration for model application[15].

Real-time Performance and Dynamic Updates: In the big data environment, data streams are constantly generated, making real-time analysis and online learning necessary requirements. How to achieve dynamic updates and real-time decision-making while maintaining high accuracy tests the comprehensive capabilities of algorithm designers.

Ⅳ. Future Development Trends and Prospects

In response to the current problems, future big data analysis methods, driven by artificial intelligence technology, will present the following development trends：

Construction of Hybrid Intelligent Models: More research will be dedicated to integrating various methods such as machine learning, deep learning, and computational intelligence. Through strategies like ensemble learning and hybrid models, higher analysis accuracy and robustness will be achieved.

Adaptive Distributed Learning Frameworks: To meet the dynamic and real-time requirements of big data, the development of adaptive distributed learning frameworks will become inevitable. By employing online learning, incremental learning, and integrated scheduling techniques, the system can maintain efficient operation in an environment where data is constantly changing.

Explainable Artificial Intelligence: As the application of artificial intelligence expands in high-risk fields, the demand for model interpretability is increasing. Future research will focus on developing methods to explain the internal mechanisms of deep learning, combining technologies such as knowledge graphs and symbolic reasoning to provide more transparent and reliable decision-making bases for big data analysis.

Edge Computing and Cloud Collaboration: With the explosive growth of data from Internet of Things devices, edge computing will be organically integrated with cloud computing to achieve local preprocessing and in-depth analysis in the cloud, thereby further reducing data transmission latency and enhancing real-time response capabilities.

Cross-domain Integration and Multi-modal Data Analysis: Big data analysis will no longer be confined to a single domain or data type. Future research will explore how to integrate various data types such as images, text, and voice through multi-modal deep learning models to achieve comprehensive analysis and intelligent decision-making in complex scenarios.

Ⅴ.conclusion

Artificial intelligence technology provides powerful methodological support for big data analysis. Through the collaborative innovation of machine learning, deep learning and computational intelligence, combined with distributed platforms such as Hadoop and Spark, scholars have achieved remarkable results in areas such as clustering, association analysis and prediction. However, algorithm efficiency, data heterogeneity and platform optimization remain unsolved problems. In the future, with the development of emerging technologies such as edge computing and quantum computing, big data analysis will move towards a higher level of intelligence and automation, driving deep changes in fields such as intelligent manufacturing and smart healthcare.

202422900232 王冉恒

References

[1]王万良,张兆娟,高楠,等.基于人工智能技术的大数据分析方法研究进展[J].计算机集成制造系统,2019,25(03):529-547.DOI:10.13196/j.cims.2019.03.001.

[2]]刘光强,干胜道."人工智能+"数字新质生产力在管理会计数字技能构建中的运用[J].财会月刊,2025,46(06):12-20.DOI:10.19641/j.cnki.42-1290/f.2025.06.002.

[3]张振博,蔡斌.大数据、云计算与人工智能技术的融合与发展[J].数字技术与应用,2025,43(01):164-166.

[4]谷庆.人工智能及大数据技术在项目管理中的应用[J].大数据时代,2024,(12):56-63.

[5]刘涛.技术驱动的电影革命:人工智能、大数据与云计算的深远影响[J].中国电影市场,2024,(12):47-51.

[6]卢星儒.大数据与智能技术在计算机网络系统中的应用[J].中国信息界,2024,(08):135-137.

[7]马川.基于移动应用的大数据存储与处理技术分析[J].电子技术,2024,53(10):44-46.

[8]赵鹏.人工智能和大数据技术在计算机网络安全防御中的运用[J].通讯世界,2024,31(09):46-48.

[9]夏商晋.大数据和人工智能背景下计算机科学与技术专业的转型发展[J].数字通信世界,2024,(09):206-208.

[10]张秀利.大数据环境下基于人工智能的网络安全态势感知技术研究[J].信息记录材料,2024,25(09):37-39.DOI:10.16009/j.cnki.cn13-1295/tq.2024.09.056.

[11]高俊.人工智能技术在智慧物流的应用研究[J].物流科技,2024,47(16):73-75+79.DOI:10.13714/j.cnki.1002-3100.2024.16.017.

[12]孙艳玲.浅析大数据背景下人工智能技术在金融领域的应用及展望[J].现代商业研究,2024,(15):49-51.

[13]周茵,赵萃.人工智能及大数据技术在数字营销中的创新应用分析[J].市场周刊,2024,37(21):100-103.

[14]王德龙.人工智能驱动的物联网大数据分析技术与未来发展趋势[J].智能物联技术,2024,56(04):1-4.

[15]郑少伟.基于人工智能技术的大数据隐私保护方法探讨[J].互联网周刊,2024,(13):53-55.