Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/104/20240904

Data mining in AI: Evolution, applications, and future directions

Zongjian Wu

The University of Queensland, Queensland, Australia

2482516799@qq.com

Abstract. This paper provides a comprehensive analysis of the evolution and impact of data mining in the field of artificial intelligence (AI), with a particular focus on its application within social and information networks. It traces the origins of AI back to the 1956 Dartmouth Conference, highlighting the subsequent advancements in technologies such as machine learning and data mining that have fueled AI's growth. The paper explores the multifaceted applications of data mining in various sectors including healthcare, transportation, and industrial manufacturing, and delves into the challenges and innovations in recommendation systems, matrix factorization, and intelligent control of autonomous vehicles in intelligent transportation systems. The study emphasizes the significance of distributed algorithms and big data processing frameworks in enhancing the efficiency and applicability of data mining techniques.

Keywords: Artificial Intelligence, Data Mining, Machine Learning, Distributed Algorithms, MapReduce Framework.

1. Introduction

The field of artificial intelligence (AI) has undergone a transformative journey since its conceptualization at the historic 1956 Dartmouth Conference. This exploration delves into the critical role of data mining in the evolution of AI, underscoring its significant impact across various sectors. At the heart of this study is an in-depth examination of the advancements in data mining technologies and their multifaceted applications, with a particular emphasis on social and information networks. Data mining, a process that involves extracting valuable information from large datasets, has become a cornerstone in the advancement of AI. It enables the uncovering of patterns, anomalies, and associations within big data, which are essential for making informed decisions and predictions. This paper focuses on how data mining has been instrumental in the development of sophisticated algorithms and models in AI. One of the key areas of discussion is the introduction of distributed algorithms in recommendation systems. These algorithms have revolutionized the way information is filtered and personalized for users in online platforms, enhancing user experience and engagement. Furthermore, the application of the MapReduce framework in matrix factorization has been pivotal in handling and processing large-scale datasets efficiently, thereby facilitating more complex data analysis and machine learning tasks. Another significant aspect of this study is the exploration of the challenges and opportunities in integrating AI into intelligent transportation systems, especially with the advent of autonomous vehicles. This integration presents a complex array of technical, ethical, and logistical challenges, yet offers immense potential for transforming urban mobility and enhancing road safety. This paper also sheds light on the

DOI: 10.54254/2755-2721/104/20240904

ethical considerations and privacy concerns surrounding data mining in AI. As AI systems become increasingly prevalent, the ethical implications of data usage, bias in algorithms, and the impact on privacy and security become critical areas of concern.

2. Research Background and Significance

2.1. Evolution and Impact of Data Mining in Artificial Intelligence

The advent of artificial intelligence dates back to the 1956 Dartmouth Conference in the United States, where the idea of "simulating human intelligence with machines" was first discussed, and the term "artificial intelligence" was introduced. Over the subsequent six decades, AI has experienced several fluctuations, evolving through different stages alongside technological advancements. Particularly since the 21st century, technologies like data mining and machine learning have triggered an explosion in AI development. Consequently, AI is regarded as a core technology and a key force in igniting the next generation of technological revolution, drawing widespread attention from various sectors. To seize the strategic high ground in this new technological revolution, major countries around the world have formulated strategies and plans for AI development. Among them, the Chinese government has placed significant emphasis on AI [2]. In July 2017, the State Council issued the "New Generation Artificial Intelligence Development Plan," aiming to make China a major global center of AI innovation by 2030.

As an integral component of AI, data mining has a broad application background. It has shown enormous potential in areas such as natural language processing, computer vision, speech recognition, and expert systems. Moreover, data mining has also attracted extensive attention in interdisciplinary fields combining security, finance, healthcare, transportation, retail, and industrial manufacturing. For example, recommendation systems can provide precise and personalized suggestions by mining features of users and items, enhancing user experience and accelerating transaction completion. By integrating apps on smartphones with camera-captured driver images, driver behavior can be mined in real-time using pattern recognition and machine learning methods. This helps in alerting drivers when they are distracted or tired, and broadcasting via vehicle networks, thus improving the intelligence and safety of the transportation system. In the security sector, data mining can quickly identify suspects from massive surveillance data, speeding up case resolution and enhancing public safety. In healthcare, data mining, combined with high-resolution medical imaging, can predict and diagnose diseases, saving lives. In industrial manufacturing, intelligent robots can provide continuous, efficient labor for specific production tasks, greatly improving the efficiency of human society.

2.2. Data Mining

Data mining, also known as knowledge discovery, aims to extract useful knowledge from large datasets. It is based on data but seeks to achieve higher-level goals, from data to information, and then to knowledge. Hence, it is considered a key step in transitioning from data to intelligence.

Common data mining techniques include classification, clustering, association analysis, regression, and recommendation systems. Classification involves "tagging" items based on their features and grouping them according to these tags [3]. Algorithms like Decision Trees, k-Nearest Neighbors (kNN), Neural Networks, and Support Vector Machines are used in areas such as churn prediction and medical diagnosis. Clustering, different from classification algorithms, groups observations into clusters based on distance metrics and does not require data labels, making it an unsupervised algorithm. Clustering algorithms, such as k-means and hierarchical clustering, are widely used in market analysis, image segmentation, and social network analysis. Association analysis seeks to find relationships among items in datasets, with the Apriori algorithm being a common method. Regression aims to model relationships between variables for prediction purposes, with linear, nonlinear, and logistic regression as common models. Recommendation systems, a significant technology and application in data mining, aim to recommend the most suitable items to users. These systems not only reduce information overload but also provide personalized services, enhancing user experience and commercial value through increased

DOI: 10.54254/2755-2721/104/20240904

click-through and conversion rates. Main algorithms in recommendation systems include Collaborative Filtering (CF), Content-Based methods (CB), and hybrid approaches.

2.3. Distributed Optimization and Processing in Big Data Analysis

The distributed optimization and processing of big data analysis provide robust support for the implementation of data mining algorithms. Distributed optimization algorithms can optimize the operations of data mining algorithms, while big data analysis frameworks facilitate more practical engineering applications.

Big data, as the foundation of AI, has attracted widespread attention in recent years. Defined by the McKinsey Global Institute, big data is characterized by vast volumes, high velocity, diverse types, and low-value density. Its rise is fueled by the explosive growth of data from internet logs, e-commerce, social networking sites, and the widespread application of computing technology. Big data often exceeds the capabilities of traditional database software tools in acquisition, storage, management, and analysis, holding immense value and potential.

Driven by the massive scale and rapid flow of big data, scalable optimization algorithms such as distributed, parallel, and online algorithms are extensively researched and applied. Distributed algorithms can alleviate computational and storage pressures on servers or computing nodes and provide a degree of privacy protection. Recently, the Alternating Direction Method of Multipliers (ADMM) has been widely used in distributed optimization and statistical learning, such as sparse logistic regression and Support Vector Machines. Parallel stochastic gradient descent algorithms are a common learning algorithm for neural networks and deep learning. Additionally, online algorithms are necessary for real-time network decisions and serial data inputs. Model Predictive Control (MPC) and other online optimization algorithms have been widely applied in real-time control scenarios like vehicle control [4].

In terms of big data processing, frameworks like Apache Hadoop, Apache Spark, and Apache Storm are extensively used in AI tasks and engineering practices. Apache Hadoop, an open-source software framework, is known for its MapReduce programming model and the Hadoop Distributed File System (HDFS). Apache Spark is particularly suited for solving effective iterative problems, while Apache Storm is apt for real-time data analysis and online machine learning. Apache Hadoop is well-suited for batch processing of large static datasets, whereas Apache Spark and Apache Storm are more appropriate for stream processing. Additionally, there exist distributed machine learning libraries such as SystemLM and MLlib, which are built upon Apache Hadoop and Apache Spark. In summary, the distributed optimization and processing of big data analysis are crucial supports for both the theoretical and practical implementation of data mining. They play a significant role in the optimization of data algorithms and in engineering practices.

2.4. Machine learning.

Machine learning aims to improve the performance of computer systems by utilizing experience, and is closely related to data mining. This section provides a brief introduction to some emerging hot technologies in machine learning. Deep learning has attracted considerable attention from academia and industry in recent years due to its ability to solve many complex problems in big data. For instance, Convolutional Neural Networks (CNNs) have achieved remarkable results in image classification and object detection; Moreover, in open-source projects, there are many libraries and frameworks for deep learning, such as TensorFlow, PyTorch, Keras, Theano, MXNet, Caffe, CNTK, Torch, DL4J, etc. As another widely followed machine learning method, Reinforcement Learning learns by interacting with an unknown environment. Q-learning, a model-free reinforcement learning algorithm, can find an optimal action selection strategy for any Markov Decision Process. Combined with deep learning, reinforcement learning has developed a series of algorithms represented by Deep Q-networks, where the Q function of action values is approximated using convolutional neural networks. These algorithms have shown outstanding performance in chess competitions, game competitions, robot control, and path planning [5]. Transfer Learning transfers knowledge learned in a current task to a related new task. In social and information networks, transfer learning will have richer application scenarios and can provide

DOI: 10.54254/2755-2721/104/20240904

customized services at a lower cost, improving the personal experience of users. In addition, some emerging learning methods have surfaced recently. Federated Learning focuses on collaborative learning while providing strong protection for data exchange security and user privacy; Meta-Learning aims to acquire the ability for small-sample learning; Automated Machine Learning focuses on learning and selecting machine learning models, features, and algorithms without human intervention; Graph Neural Networks are dedicated to running neural networks on graph data structures; and so on. Figure 1 provides an overview of some machine learning algorithms, noting that some algorithms and techniques are mentioned again due to the intersection of data mining and machine learning.

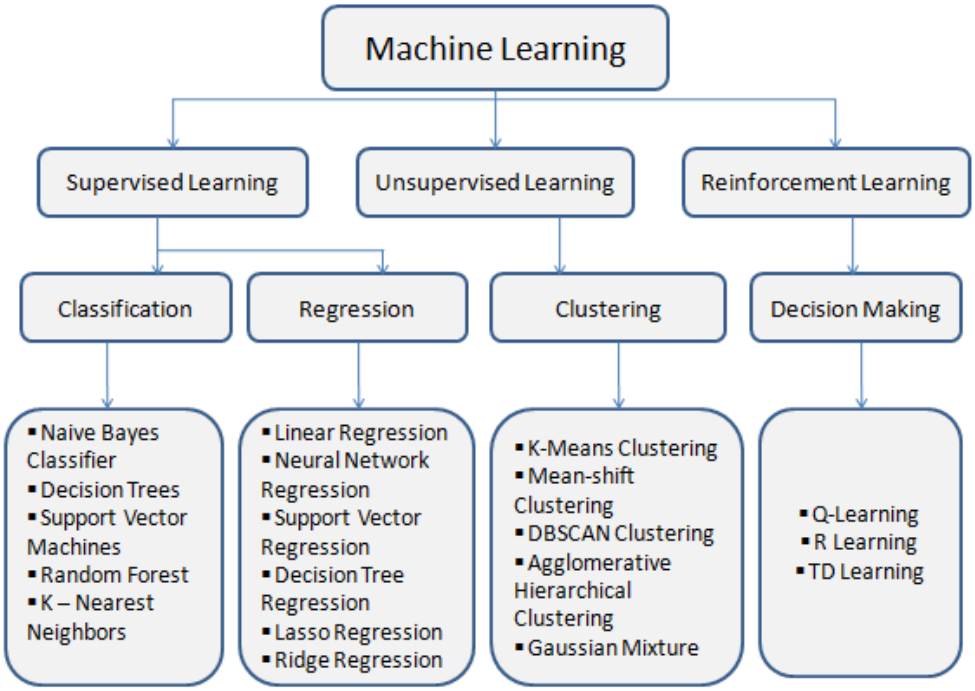

Figure 1. Overview of Machine Learning Algorithms

2.5. Challenges of Data Mining in Social and Information Networks

Data mining faces several challenges when applied in social and information networks. In social networks, recommendation systems can not only obtain traditional user-item rating matrices but also a large amount of additional information such as content, context, and friend trust. However, similarity constraints introduced to consider additional information pose difficulties for the distributed implementation of recommendation systems. Existing MapReduce frameworks for matrix factorization in information networks often incur significant startup overhead due to frequent generation of MapReduce jobs during iteration, and the structure of matrix factorization is not fully exploited. In intelligent transportation networks where unmanned vehicles coexist with human-driven vehicles, the complex behavior of human drivers, such as distraction and fatigue, is difficult for unmanned vehicles to perceive, posing significant safety risks to both unmanned vehicles and the entire transportation network [6]. In information networks, decision-makers often need to immunize and control information dissemination without knowing the initial source of propagation, making it crucial to effectively allocate limited resources to control the worst-case network spread.

DOI: 10.54254/2755-2721/104/20240904

3. Advancements in Data Mining within Social and Information Networks

3.1. Distributed Algorithms for Recommendation Systems in Social Networks

Recommendation systems are widely applied in social networks, enhancing user experience and promoting business conversions. For instance, in movie recommendation systems, their schematic is illustrated in Figure 2. The design of distributed algorithms for recommendation systems has emerged as a research hotspot due to requirements for data scale and content freshness. Distributed algorithms offer enhanced robustness and a degree of privacy protection compared to centralized models. The Alternating Direction Method of Multipliers (ADMM) has garnered significant attention and is extensively used in optimization and machine learning, sparse low-rank decomposition, resource management, multi-cell cooperative beamforming, and average consensus problems. Specifically for matrix factorization techniques often used in recommendation systems, ADMM is frequently employed in the design of distributed matrix factorization. Yu et al. proposed a distributed stochastic ADMM using data splitting strategies and stochastic mechanisms to solve large-scale matrix factorization problems. Du et al. adopted ADMM to separate the objective minimization from non-negative constraints to solve non-negative matrix factorization problems. Cai et al. utilized ADMM to solve matrix factorization with maximum norm constraints, providing proofs for convergence to optimal solutions for convex problems. Mardani et al. introduced a decentralized algorithm based on ADMM for nuclear norm-based matrix factorization [7].

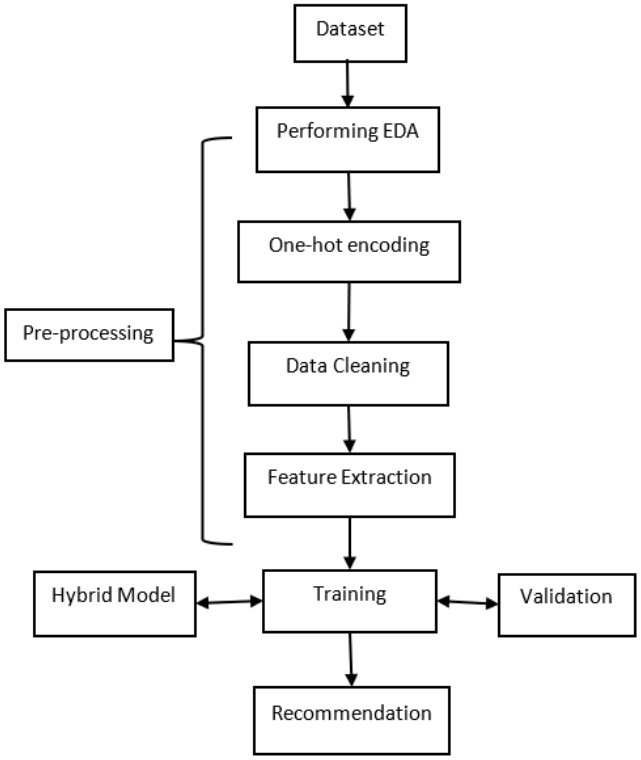

Figure 2. Movie Recommendation System Schematic

3.2. The MapReduce Framework for Matrix Factorization in Information Networks

In information networks, matrix factorization, which decomposes a target matrix into the product of two low-rank matrices, is extensively used for dimension reduction and matrix completion. It finds applications in collaborative filtering, recommendation systems, network distance prediction, and wireless sensor localization. The MapReduce framework is widely adopted in both industry and academia for big data analytics, but it incurs high startup overhead with each iterative MapReduce job,

DOI: 10.54254/2755-2721/104/20240904

making it unsuitable for typical matrix factorization algorithms. On one hand, there have been efforts to tailor the MapReduce framework for matrix factorization: Gemulla et al. introduced a Hadoop implementation using distributed SGD for matrix factorization, employing map-only jobs in each SGD phase to avoid extensive data shuffling in the reduce phase. Schelter et al. implemented ALS by assigning a map-only job to each factor matrix and server through a series of broadcast-joins. On the other hand, neither iterative MapReduce implementations like Haloop and Twister, nor general-purpose machine learning platforms like MLI and SystemML, fully exploit the inherent structure of matrix factorization to modify the MapReduce framework.

3.3. Intelligent Control of Autonomous Vehicles in Intelligent Transportation Networks

In intelligent transportation networks, autonomous vehicles will become key participants. Coexisting with human-driven vehicles, the intelligent control of autonomous vehicles and their interaction with human drivers are increasingly important. Current research on autonomous vehicle control mainly focuses on platooning, lane-changing, or assisted driving, often overlooking interactions with human-driven vehicles. Although some studies address shared road strategies at intersections and emergency warnings, they do not fully consider complex human driver behaviors like distraction. Recent studies have started to incorporate the role of human drivers in autonomous driving, such as Lefèvre et al.'s framework that learns from human demonstrations for autonomous driving. Tehrani et al. compared lane-changing maneuvers between human drivers and computer-generated actions on highways. Analyzing human lane-changing data, Do et al. proposed a two-stage lane-changing model to mimic human drivers [8].

4. Conclusion

In conclusion, data mining has emerged as a cornerstone technology in the realm of artificial intelligence, driving significant advancements and applications across various domains. The evolution of data mining techniques, particularly in social and information networks, has demonstrated its versatility and capacity to address complex challenges. While the development of distributed algorithms and the MapReduce framework have catalyzed the efficiency and scalability of data mining processes, there remain areas that require further exploration, such as the integration of additional information in recommendation systems and the intelligent control of autonomous vehicles. This paper underscores the importance of continuous innovation and research in data mining to harness its full potential in the ever-evolving landscape of artificial intelligence.

References

[1] Ageed, Zainab Salih, et al. "Comprehensive survey of big data mining approaches in cloud systems." Qubahan Academic Journal 1.2 (2021): 29-38.

[2] Stephany, Fabian, et al. "The CoRisk-Index: A data-mining approach to identify industry-specific risk assessments related to COVID-19 in real-time." arXiv preprint arXiv:2003.12432 (2020).

[3] Van Nguyen, Truong, et al. "Predicting customer demand for remanufactured products: A data-mining approach." European Journal of Operational Research 281.3 (2020): 543-558.

[4] Greener, Joe G., et al. "A guide to machine learning for biologists." Nature Reviews Molecular Cell Biology 23.1 (2022): 40-55.

[5] Jo, Taeho. "Machine learning foundations." Supervised, Unsupervised, and Advanced Learning. Cham: Springer International Publishing (2021).

[6] Mahesh, Batta. "Machine learning algorithms-a review." International Journal of Science and Research (IJSR).[Internet] 9.1 (2020): 381-386.

[7] Janiesch, Christian, Patrick Zschech, and Kai Heinrich. "Machine learning and deep learning." Electronic Markets 31.3 (2021): 685-695.

[8] Murphy, Kevin P. Probabilistic machine learning: an introduction. MIT press, 2022.