Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 3, 2024

Research on data preprocessing method for artificial intelligence algorithm based on user online behavior

Zhiyuan Liu

Chongqing University of Arts and Sciences, Chongqing 402160, China
owenkeswick2024@gmail.com

Abstract: This paper delves into the significance and efficacy of data preprocessing techniques specifically tailored for user online behavior data in enhancing the performance of artificial intelligence algorithms. Through a comprehensive literature review, we identify gaps and opportunities in current methodologies, setting the stage for the development of a novel data preprocessing approach. This proposed method is meticulously designed to handle the complexities and nuances of user online behavior data. We conduct a series of case studies and rigorous empirical analyses to evaluate the effectiveness of our approach. The results clearly demonstrate that our method substantially improves data quality by effectively reducing noise and eliminating irrelevant information, which, in turn, enhances the overall performance of the AI algorithms. The paper concludes with a discussion on the limitations of the current study and provides insightful directions for future research in the field. This includes potential refinements to the preprocessing technique and its application to other types of behavioral data in different AI domains.

Keywords: User online behavior; Data preprocessing; Artificial intelligence algorithms; Data quality.

1. Introduction

With the popularity of the Internet and the rapid development of information technology, user online behavior data has become a valuable resource with wide applications in various fields. Particularly in the field of artificial intelligence, mining and analyzing user online behavior data has become an important driving force for algorithm development and application. However, user online behavior data often presents challenges due to its complexity, diversity, and uncertainty. Data preprocessing, as an indispensable part of the data mining and machine learning process, plays a crucial role in ensuring data quality, reducing noise interference, and improving algorithm performance. However, there is a relative lack of preprocessing methods tailored for user online behavior data, and existing methods often fail to fully exploit the potential information and value of the data, limiting the effectiveness of algorithms. Therefore, this paper aims to explore data preprocessing methods based on user online behavior to improve data quality, reduce noise interference, and optimize the performance of artificial intelligence algorithms. Through in-depth analysis of relevant theories and empirical research, this paper will propose an effective data preprocessing method and evaluate its performance on real-world datasets. This will provide important theoretical and practical guidance for further exploring the potential of user online behavior data and advancing the development of artificial intelligence algorithms [1].

2. Artificial Intelligence Algorithms' Current Applications and Shortcomings in Data Preprocessing

2.1. Overview of Artificial Intelligence Algorithms' Applications in Data Preprocessing

With the rapid evolution of artificial intelligence (AI) technologies, a diverse array of AI algorithms have been extensively employed in the domain of data preprocessing. Key among these are deep learning, machine learning, and data mining algorithms, which serve as foundational tools for transforming raw data into a refined format suitable for analytical models. Deep Learning Applications: Deep learning algorithms, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are pivotal in preprocessing tasks for various data types such as images, text, and speech. CNNs are adept at extracting high-level, abstract features from image data through sophisticated operations like convolution and pooling layers. This capability makes them ideal for tasks requiring detailed visual understanding, thereby enhancing the quality of data fed into subsequent analytical processes. On the other hand, RNNs are designed to handle sequential data, making them effective in recognizing patterns over time and capturing temporal correlations and sequence-specific features. This is particularly beneficial in fields like natural language processing and time-series analysis. Machine Learning Applications: Algorithms such as Support Vector Machines (SVMs) and Random Forests play crucial roles in classification, regression, and clustering tasks. These machine learning models are trained on preprocessed data to perform predictions and classifications, which supports decision-making processes across various applications. The adaptability and robustness of these algorithms allow them to be applied in scenarios ranging from spam detection to financial forecasting. Data Mining Applications: Data mining techniques such as association rule mining, clustering analysis, and anomaly detection are instrumental in uncovering underlying patterns and rules within large datasets. These techniques facilitate the identification of significant insights that can influence the direction of further data analysis and preprocessing strategies. Despite these advancements, the application of AI algorithms in data preprocessing faces several challenges. The complexity of some algorithms makes them difficult to implement and optimize without substantial computational resources and

expert knowledge. Additionally, issues related to data quality, such as missing values and noise, can adversely affect the performance of AI algorithms. There is also a need for algorithms to be more adaptable to specific domains, as current one-size-fits-all approaches may not always yield optimal results in specialized fields. In summary, while the integration of AI algorithms into data preprocessing has achieved notable successes, it also presents a landscape ripe for further research and development. Improvements in algorithmic design, along with enhanced understanding and handling of domain-specific data characteristics, are essential for advancing this field[2].

2.2. Data Preprocessing Methods for User Online Behavior

Preprocessing user online behavior data is a crucial step in the application of artificial intelligence algorithms. When dealing with user online behavior data, we face many challenges due to its specificity and complexity. Firstly, data cleaning and denoising are indispensable steps. The original data often contains problems such as duplicates, missing values, or outliers. By using techniques such as filtering and outlier detection, we can effectively remove incomplete or interfering data, improving the accuracy and reliability of the data. Secondly, feature extraction and selection are also crucial. User online behavior data typically contains numerous features, and we need methods such as Principal Component Analysis (PCA) and information gain to reduce the dimensionality of the data while retaining the most representative features for subsequent modeling and analysis. Additionally, smoothing and processing time series features in user online behavior data using methods such as time windows and moving averages are essential for capturing the time correlation and periodicity of the data. Furthermore, data standardization and normalization are indispensable steps to eliminate the dimensional differences between different features, ensuring the stability and convergence of the model. Simultaneously, anomaly detection and processing are also important. User online behavior data may contain abnormal behaviors or accesses that need to be identified and processed through methods such as clustering and statistical analysis to ensure the quality and reliability of the data. Finally, data fusion and integration are also critical. For user online behavior data from different sources or formats, we need to perform data fusion and integration to unify the data format and structure for subsequent analysis and modeling. In summary, data preprocessing methods for user online behavior involve multiple aspects such as data cleaning, feature extraction, time series processing, data standardization, anomaly detection, and data fusion. By comprehensively applying these methods, we can effectively improve the quality and usability of user online behavior data, providing a reliable foundation for subsequent data analysis and modeling [3].

2.3. Shortcomings in the Application of Artificial Intelligence Algorithms in Data Preprocessing

Although artificial intelligence algorithms play an important role in data preprocessing, there are still some shortcomings. Firstly, algorithm complexity is a significant issue. Some advanced artificial intelligence algorithms, such as deep learning models, require a large amount of computational resources and expertise to implement and adjust. This may be a barrier for some researchers and practitioners, limiting the widespread use and application of the algorithms. Secondly, data quality issues are another key factor affecting the effectiveness of algorithms. User online behavior data often contains noise, missing values, and incomplete data, which may affect the accuracy and reliability of the algorithms. Current artificial intelligence algorithms still have limitations in addressing these issues, requiring more sophisticated and effective methods to solve them. Additionally, the lack of flexibility and applicability of general artificial intelligence algorithms for specific domain data preprocessing needs is also a challenge. Data in different domains have their own characteristics and patterns, requiring tailored design and adjustment of preprocessing methods. Currently, there is a relative lack of customized preprocessing methods for user online behavior data, requiring further research and exploration. Finally, the interpretability and transparency of algorithms are also a challenge. Some complex artificial intelligence algorithms, such as deep neural networks, are often considered black box models, making it difficult to explain their decision-making processes and internal mechanisms. This limits the application of the algorithms in scenarios where interpretability is required, such as the medical and financial fields. In conclusion, although artificial intelligence algorithms play an important role in data preprocessing, there are still many challenges and shortcomings. Future research needs to focus on addressing these issues, improving the performance and applicability of the algorithms, and promoting further development and application of artificial intelligence technology in the field of data preprocessing.

3. Basic Concepts and Theoretical Framework

3.1. The Basic Concepts and Importance of Data Preprocessing

Data preprocessing, as a preliminary step in data mining and machine learning, plays a crucial role. Its basic concepts encompass operations such as cleaning, transforming, integrating, and normalizing raw data to make it suitable for subsequent analysis and modeling processes. During data preprocessing, the data we deal with often comes from various sources, formats, qualities, and characteristics. These raw data typically contain issues like noise, missing values, outliers, and data inconsistency, which may lead to inaccuracies and biases if directly applied to algorithm models. Therefore, the importance of data preprocessing is self-evident. Firstly, data preprocessing can enhance the quality and usability of data. By cleaning data, handling missing values, and outliers, we can reduce the impact of noise, improve data accuracy and reliability, laying a reliable foundation for subsequent analysis and modeling. Secondly, data preprocessing can reduce the complexity and computational costs of models. Through operations such as feature selection and dimensionality reduction, we can decrease data dimensions and complexity, improving model training efficiency, prediction performance, and reducing the risk of overfitting. Additionally, data preprocessing can enhance the generalization ability and stability of models. By standardizing, normalizing data, we can eliminate dimensional differences between different features, making models more stable and reliable, and improving their adaptability and prediction accuracy on new data. In summary,

data preprocessing is an indispensable part of the data analysis and modeling process. Its importance lies in improving data quality, reducing model complexity, and enhancing model generalization ability. Effective data preprocessing methods and techniques can improve algorithm efficiency and accuracy, providing a solid foundation for data-driven decision-making and applications [4].

3.2. Data Preprocessing Process in Artificial Intelligence Algorithms

The data preprocessing process within artificial intelligence (AI) algorithms is a critical phase aimed at ensuring data quality and extracting meaningful features. This process encompasses several key steps, each contributing to the overall effectiveness of the algorithm. First and foremost, data cleaning serves as the foundational step in data preprocessing. It involves the identification and removal of errors, incompleteness, inconsistencies, and duplicates within the dataset. By addressing these issues, data integrity and consistency are upheld, laying a robust groundwork for subsequent analysis. Following data cleaning, the stage of feature selection and extraction comes into play. This step focuses on identifying the most relevant and informative features or engineering new ones through sophisticated techniques. By reducing data dimensions and enhancing the efficiency of model training, feature selection and extraction contribute significantly to improving the algorithm's predictive performance while mitigating the risk of the curse of dimensionality. Subsequently, data transformation and normalization play pivotal roles in preparing the data for model consumption. Techniques such as normalization, standardization, and discretization are employed to ensure that the data is presented in formats and ranges compatible with the model's requirements. By eliminating disparities between different features, these methods facilitate the model's ability to discern patterns and rules within the data. Furthermore, dataset division is an essential aspect of the data preprocessing pipeline. This involves partitioning the dataset into subsets, such as training, validation, and testing sets. By doing so, the model's performance and generalization capacity can be effectively evaluated, while simultaneously guarding against overfitting during the training phase. In instances where the dataset is limited in size or exhibits imbalanced samples, data augmentation techniques can be employed to bolster the dataset. Methods such as random rotation, flipping, cropping, among others, are utilized to generate synthetic samples, thereby enhancing the model's generalization capabilities and robustness. In conclusion, the data preprocessing process within AI algorithms encompasses a series of indispensable steps, including data cleaning, feature selection and extraction, data transformation and normalization, dataset division, and potentially data augmentation. Through meticulous execution of these steps, the performance and generalization ability of the model are enhanced, thereby enabling its more effective application to real-world problems. By laying a solid foundation through data preprocessing, AI algorithms can derive actionable insights and solutions that drive impactful outcomes across various domains and industries [5].

3.3. Exploring the Characteristics and Challenges of User Online Behavior Data

User online behavior data presents a multitude of unique characteristics and challenges, necessitating a deep understanding for effective data preprocessing. Firstly, the diversity of user online behavior data is striking. With the widespread use of the Internet and the proliferation of digital technologies, online user behaviors encompass a wide array of activities, including web browsing, information retrieval, e-commerce transactions, social interactions, and more. This diversity results in a complex and heterogeneous landscape of user online behavior data structures, mandating the adoption of tailored processing methods to accommodate the varying types of data. Secondly, user online behavior data exhibits characteristics of high dimensionality and large-scale. The exponential growth of Internet usage has led to an unprecedented volume of data generated by users, including clickstreams, browsing histories, search queries, and other forms of interaction. These datasets often feature a multitude of dimensions and variables, necessitating the utilization of efficient processing techniques to reduce data dimensionality and extract meaningful features. Moreover, the temporal and dynamic nature of user online behavior data poses significant challenges [6]. Online user behaviors exhibit temporal patterns and dynamics, with behavior varying across different time periods. Consequently, processing user online behavior data requires methodologies such as time series analysis and dynamic modeling to capture temporal features and fluctuations within the data effectively. Furthermore, user online behavior data raises concerns regarding privacy and security. Given that online behavior often involves the disclosure of personal information and sensitive data, such as individual preferences and consumption habits, ensuring adequate privacy protection and security measures is imperative to safeguard user privacy rights and data integrity. In summary, user online behavior data is characterized by its diversity, high dimensionality, temporality, and privacy and security considerations, all of which present significant challenges for data preprocessing. Addressing these challenges effectively entails the adoption of diverse data processing methods and technologies to enhance data quality, reduce dimensionality, capture temporal dynamics, and safeguard user privacy. By doing so, a solid foundation can be established for subsequent data analysis and modeling, facilitating informed decision-making and improving user experiences in online environments.

4. Data Preprocessing Methods

4.1. Key Steps in Data Preprocessing for User Online Behavior

Data preprocessing plays a pivotal role in ensuring the quality and effectiveness of user online behavior data analysis. Understanding the key steps involved in this process is essential for researchers and practitioners alike. First and foremost, data cleaning serves as the foundational step in data preprocessing. Its primary objective is to ensure data integrity and accuracy by addressing issues such as duplicates, missing values, and outliers. By removing or correcting these inconsistencies, the reliability of the dataset is significantly enhanced, laying a robust foundation for subsequent analysis. Following data cleaning, feature extraction and selection become crucial steps in the preprocessing pipeline. Feature engineering techniques allow for the identification of relevant features that best represent the underlying patterns in the data. This process aids in reducing the dimensionality of the dataset while simultaneously improving the efficiency and predictive performance of the models trained on it. Given the temporal

nature of user online behavior data, time series processing emerges as a necessary component of preprocessing. Techniques such as time window analysis and moving averages enable the capture of temporal patterns and trends within the data. By analyzing user behavior over specific time intervals, insights into evolving trends and patterns can be gleaned, facilitating more accurate analysis and prediction. Data standardization and normalization are imperative for ensuring consistency and comparability across different features of the dataset. These techniques aim to eliminate variations in scale and magnitude, thereby enhancing model stability and convergence. Common normalization methods like Z-score normalization and min-max scaling are widely employed to achieve this objective. Furthermore, anomaly detection and handling are essential for identifying and mitigating outliers or abnormal behavior within the dataset. By identifying and addressing these anomalies, the overall quality and reliability of the data are improved, ensuring that the models are trained on representative and trustworthy data. Finally, if user online behavior data originates from diverse sources or formats, data fusion and integration techniques are employed to consolidate the disparate data into a unified format and structure. This process enables seamless integration and compatibility, facilitating more comprehensive analysis and modeling. In summary, by adhering to these key steps in data preprocessing—data cleaning, feature extraction and selection, time series processing, data standardization and normalization, anomaly detection and handling, and data fusion and integration—the quality, consistency, and usability of user online behavior data can be significantly enhanced. This, in turn, provides a solid foundation for more accurate and insightful data analysis and modeling, ultimately driving informed decision-making and improving user experiences in online platforms [7].

4.2. Common Data Preprocessing Techniques and Their Applicability in this Field

In the domain of preprocessing user online behavior data, leveraging common preprocessing techniques is essential for effective data preparation. These techniques serve as foundational steps to ensure the integrity, accuracy, and usability of the data for subsequent analysis and modeling. Firstly, data cleaning and processing stand as the bedrock of data preprocessing. This critical task involves the identification and removal of duplicates, handling missing values, and filtering out anomalies. By addressing these issues, data integrity and accuracy are preserved, laying a solid foundation for reliable analysis and modeling. Secondly, feature extraction and selection are pivotal in reducing data dimensionality and enhancing model efficiency and performance. Techniques such as principal component analysis (PCA), information gain, and variance thresholding aid in identifying the most relevant features, thus improving model generalization and predictive capabilities. By selecting the most informative features, unnecessary noise and redundant information are minimized, leading to more effective and efficient models. Moreover, data standardization and normalization are fundamental techniques to address dimensional disparities between features. Standardizing and normalizing the data eliminate scale differences, ensuring that each feature contributes equally to the modeling process. Common methods like Z-score normalization and min-max scaling are widely employed to achieve this goal, enhancing model stability and convergence. Given the temporal nature of user online behavior data, time series processing assumes critical importance. Techniques such as time window analysis and moving averages are employed to capture temporal patterns and trends within the data. By analyzing data over specific time intervals, insights into user behavior dynamics and trends can be gleaned, facilitating more informed decision-making and analysis. Furthermore, anomaly detection and handling are indispensable for identifying and mitigating outliers or abnormal behavior within the data. By detecting and addressing anomalies, data quality and reliability are improved, ensuring that the models are trained on robust and representative data. Lastly, for data originating from diverse sources or formats, data fusion and integration techniques are employed to consolidate disparate sources into a unified format and structure. By integrating data from multiple sources, a more comprehensive and holistic understanding of user behavior can be achieved, enabling more accurate analysis and modeling. In summary, these common data preprocessing techniques hold significant applicability in the preprocessing of user online behavior data. By employing these techniques, data quality is improved, model complexity is reduced, and model performance is enhanced, ultimately facilitating more accurate and reliable data-driven insights and decisions [8].

4.3. Exploration of Emerging Data Preprocessing Methods and Technologies

In the field of preprocessing user online behavior data, emerging data preprocessing methods and technologies continually emerge, offering new insights and solutions to challenges faced by existing techniques. Some emerging data preprocessing methods and technologies include: Firstly, deep learning-based preprocessing methods are gaining attention. Deep learning models excel in feature learning, automatically capturing high-level representations of data. Techniques such as autoencoders, convolutional neural networks offer potential in end-to-end data preprocessing and feature learning, improving model performance and generalization. Secondly, image processing-based preprocessing methods are becoming prominent. User online behavior data often includes image information like webpage screenshots, user avatars, etc. Leveraging image processing techniques for preprocessing and feature extraction, such as image segmentation, feature extraction, object recognition, enriches data dimensionality and information. Additionally, natural language processing (NLP) techniques are gradually rising in preprocessing user online behavior data. With abundant textual information in user online behavior data such as search records, comments, social media content, NLP techniques enable preprocessing and feature extraction tasks like text tokenization, word embeddings, sentiment analysis, revealing richer semantic information. Moreover, reinforcement learning-based preprocessing methods are garnering attention. Reinforcement learning, through interaction with the environment, can automatically adjust parameters and strategies in the preprocessing process, achieving adaptive optimization and improving model performance and robustness. In conclusion, emerging data preprocessing methods and technologies offer new perspectives and solutions for preprocessing user online behavior data. Explorations in deep learning, image processing, natural language processing, and reinforcement learning provide opportunities and challenges in improving preprocessing efficiency and model performance.

5. Conclusion

In the research on preprocessing user online behavior data, we have delved into the importance of data preprocessing, current commonly used preprocessing methods and technologies, as well as emerging exploration directions. Through the analysis of key steps such as data cleaning, feature extraction, data standardization, time series processing, anomaly detection, and data fusion in preprocessing user online behavior data, we recognize the crucial role of data preprocessing in improving data quality, reducing model complexity, and enhancing model performance. Currently common preprocessing methods include data cleaning and processing, feature extraction and selection, data standardization and normalization, time series processing, anomaly detection and handling, and data fusion and integration. These methods play important roles in improving data quality, reducing dimensionality, and enhancing model robustness. Additionally, emerging preprocessing methods and technologies, such as exploration in deep learning, image processing, natural language processing, and reinforcement learning, offer new perspectives and solutions for preprocessing user online behavior data. Despite making certain research progress, there are still many challenges and unresolved issues, such as ensuring data quality, model interpretability, and privacy security protection. Therefore, future research needs to further explore new data preprocessing methods and technologies, improve the efficiency and quality of data processing, advance the development of preprocessing user online behavior data field, and provide more reliable support for data-driven intelligent applications.

References

[1] Chakoory O, Barra V ,Rochette E , et al.DeepMPTB: a vaginal microbiome-based deep neural network as artificial intelligence strategy for efficient preterm birth prediction. [J].Biomarker research,2024,12(1):25-25.

[2] Khalifa D A, Noora F ,Murat K .Artificial Intelligence and Cyber Defense System for Banking Industry: A Qualitative Study of AI Applications and Challenges[J].Cybernetics and Systems, 2024, 55(2):302-330.

[3] Saad A S, Shayea I ,Ahmed S M N .Artificial intelligence linear regression model for mobility robustness optimization algorithm in 5G cellular networks[J].Alexandria Engineering Journal, 2024, 89125-148.

[4] Liang Y, Li G ,Xu M , et al.An intelligent control method based on artificial neural network for numerical flight simulation of the basic finner projectile with pitching maneuver[J].Defence Technology, 2024,32663-674.

[5] Simona R Joshua E .Blockchain for Artificial Intelligence (AI): enhancing compliance with the EU AI Act through distributed ledger technology. Acybersecurity perspective[J]. International Cybersecurity Law Review, 2024, 5(1):1-20.

[6] Dianshuai D, Hongliang F .Design and use of a wireless temperature measurement network system integrating artificial intelligence and blockchain in electrical power engineering. [J]. PloS one, 2024,19(1):e0296398-e0296398.

[7] Gupta S, Shetty S ,Natarajan S , et al.A comparative evaluation of concordance and speed between smartphone app-based and artificial intelligence web-based cephalometric tracing software with the manual tracing method: A cross-sectional study. [J].Journal of clinical and experimental dentistry, 2024, 16(1):e11-e17.

[8] Juanjuan R ,Salwa S S .Green urban logistics path planning design based on physical network system in the context of artificial intelligence[J].The Journal of Supercomputing, 2023, 80(7):9140-9161.