Proceedings of CONF-SEML 2026 Symposium: Multimodal Data Acquisition: Applications in Physiological and Behavioral Research
DOI: 10.54254/2755-2721/2026.AD31997

A Review of Data Preprocessing Techniques in Big Data Analysis

Zikun Peng

New Channel Jinqiu College, Beijing, China
18701318560@163.com

Abstract. With the full arrival of the big data era, data has gradually become a core strategic asset for scientific decision-making across industries. However, raw data often suffers from issues such as missing values, noise, inconsistencies, and redundancy due to diverse sources and inconsistent formats, which directly impair the quality and credibility of data analysis. As a critical component of the big data analysis process, data preprocessing plays a vital role in enhancing data quality and standardizing data formats. The effectiveness of preprocessing directly determines the accuracy and reliability of subsequent modeling and analysis. This paper systematically reviews and summarizes the core technologies involved in data preprocessing for big data analysis. Based on an extensive literature review and inductive analysis methods, it focuses on analyzing the fundamental principles and typical processing methods of key preprocessing steps, including data cleaning, data integration, data transformation, and data reduction. By examining practical applications in industries such as financial risk control, medical diagnosis, and e-commerce, the paper explores the real-world scenarios and outcomes of these technologies. Additionally, it delves into major challenges in current data preprocessing, including the complexity of data quality assessment, computational efficiency issues in high-dimensional data processing, and the growing importance of data privacy and security protection. The study concludes that efficient and intelligent data preprocessing is a prerequisite for fully unlocking the value of big data. Future research directions will increasingly focus on developing and optimizing automated, adaptive preprocessing technologies and integrated frameworks.

Keywords: big data, data preprocessing, data cleaning, data integration, data mining

1. Introduction

Nowadays, with the rapid development of information technology, the global data scale is growing explosively, and big data analysis technology has become the key to extracting high-value information from massive and diverse data. However, the principle of "garbage in, garbage out" still applies in the big data environment. Low-quality and unprocessed raw data will cause analysis models to produce deviations or even draw wrong conclusions, affecting the effectiveness of decision-making. Therefore, data preprocessing, as a basic link to improve data quality and enhance the credibility and robustness of analysis results, is increasingly important in strategic positions.

This paper takes data preprocessing technology in the big data environment as the research object and conducts a systematic sorting and review. The research adopts the literature review method to comprehensively summarize and comment on the mainstream key data preprocessing technologies and their implementation mechanisms, and discusses around three core issues: (1) What are the key technologies, core principles and typical methods of data preprocessing? (2) How are these technologies applied in different industries to cope with actual data processing challenges? (3) What bottlenecks and challenges are currently faced by data preprocessing technologies, and what are the future development directions and trends? In-depth exploration of this theme is helpful to promote the practical application of big data analysis technology and has important theoretical and practical value for improving the level of data-driven decision-making in various fields.

2. The importance of data preprocessing in big data analysis

Data preprocessing is the "preliminary core link" in the entire process of big data analysis, which performs "quality correction and value extraction" on raw data. Raw data has problems such as missing values and noise. Through operations like cleaning, preprocessing converts it into usable samples, covering the "data access—data governance—data output" chain, and serving as a necessary bridge connecting "data production" and "data value mining".

The "flaws" in raw data can lead to the "Garbage In, Garbage Out" issue, as exemplified in the retail and medical fields. The Big Data Analysis Technology White Paper (2024) shows that improving the quality of the preprocessing link can enhance the prediction accuracy of subsequent machine learning models and the F1 score of classification tasks, confirming that it plays a decisive role in ensuring the reliability of analysis results [1].

In data mining scenarios, preprocessing is the prerequisite for "effective pattern discovery"; in machine learning scenarios, it is the core of "ensuring the generalization ability of models". In short, without high-quality preprocessing, the results of data mining and machine learning have no actual business value.

3. A review of key technologies in data preprocessing

3.1. Data cleaning

The data cleaning includes missing value processing, noise identification and smoothing techniques, anomaly detection and processing.

Missing values are a common issue in raw data, with diverse causes. The selection of processing methods depends on the proportion and type of missing values, primarily including deletion and imputation. Deletion is suitable for samples or features with less than 5% missing values, but may lead to data loss. Imputation is the mainstream approach, encompassing statistical imputation (mean, median, and mode imputation), model-based imputation (K-nearest neighbors and regression imputation), and advanced imputation (multiple imputation).

Noise data refers to meaningless, erroneous, or abnormal values that interfere with analysis. It is characterized by being difficult to process, diverse in form, and affecting results. Methods for identifying noise data include statistical methods, clustering methods, and modeling methods; smoothing techniques include moving average smoothing, regression smoothing, and bin smoothing.

Anomalies refer to data points that significantly deviate from normal distribution patterns or disrupt analytical results, requiring statistical methods or domain knowledge for identification and processing. Detection approaches include distance-based (using LOF to calculate distances between samples and their neighbors to identify outliers), density-based (applying DBSCAN to flag low-density samples as anomalies), statistical methods (employing interquartile range or Grubbs test), and machine learning techniques (using isolated forests to rapidly identify outliers). Handling strategies involve removing business-irrelevant anomalies (e.g., "user age = 200"); correcting traceable anomalies (e.g., replacing "order amount = 0" with actual payment values); flagging business-critical anomalies (e.g., "high-value customer's large single transaction" as "special cases" for analysis); and binning outliers to prevent interference with overall analysis.

3.2. Data integration

Multi-source data fusion is the process of integrating data from different systems and formats into a unified dataset, with entity recognition as its core. This involves identifying records in different data sources that refer to the same entity, such as different systems corresponding to the same user. Entity recognition methods include: (1) attribute similarity matching, where entities are identified by exceeding a threshold; (2) rule-based matching, where business rules are combined to construct matching rules; (3) machine learning-based entity linking, where pre-trained language models convert entity attributes into vector matches; and (4) knowledge graph matching, where industry knowledge graphs are constructed to assist in matching.

Data conflicts occur when the same entity's attribute values differ across data sources. These conflicts can be detected through the "attribute-entity" mapping table. Resolution mechanisms include authoritative data source priority, timestamp priority, majority voting, business rule validation, and manual review.

In enterprise data applications, data integration relies on "data warehouses" and "ETL processes," with "dimensional modeling" at its core. This process divides data into "dimensional tables" and "fact tables" to integrate multi-source data. A data warehouse is a system designed for data storage and analysis, supporting business decision-making by consolidating multi-source data and facilitating historical data analysis and reporting. ETL integration technologies include extraction (using CDC technology to capture incremental data in real-time or periodic tasks to collect full datasets), transformation (performing data cleaning and other preprocessing steps), loading (loading transformed data into the warehouse's dimensional and fact tables), and scheduling (using scheduling tools to execute and monitor ETL processes periodically).

3.3. Data conversion

Data transformation requires initial standardization and normalization, a critical step in data preprocessing that enhances data quality and model performance. The core objective is to eliminate dimensionality differences among features. Standardization methods include minimum-maximum normalization (scaling data to the [0,1] range, suitable for datasets with known distributions and no outliers) and absolute maximum normalization (scaling data to the [-1,1] range, applicable to datasets containing both positive and negative values). Normalization methods comprise Z-score normalization (transforming data into a distribution with mean 0 and variance 1, ideal for near-normal datasets) and robust normalization (replacing mean with median and standard deviation with interquartile range, effective for datasets with outliers).

Furthermore, data discretization and aggregation techniques are two critical approaches in data transformation. Discretization methods include equal-width discretization (dividing intervals by fixed width, suitable for uniform data distribution), equal-frequency discretization (creating intervals with equal sample sizes, ideal for non-uniform distributions), cluster discretization (segmenting data into clusters using clustering algorithms, effective for complex distributions), and decision tree discretization (applying decision tree models to identify the most information-gain-enhancing partition points, particularly useful in supervised learning scenarios). Aggregation techniques encompass temporal aggregation (e.g., aggregating "daily sales" into "weekly" or "monthly" sales to analyze trends), spatial aggregation (e.g., consolidating "store sales" into "regional sales" to assess market performance), and dimensional aggregation (e.g., merging "user-product" transaction data into "user-category" data to analyze consumer preferences).

Meanwhile, feature construction and dimensionality reduction are crucial components of data preprocessing and feature engineering. Feature construction involves generating new features from existing data to address practical problems and enhance model performance, with the core focus on information reorganization. Dimensionality reduction employs mathematical methods to transform high-dimensional features into low-dimensional ones, primarily aimed at simplifying model complexity. Common feature construction methods include business logic-based feature generation (derived from operational scenarios), statistical feature construction (calculating feature statistics), interaction feature construction (creating multiple feature interactions), and text feature construction (segmenting text data and generating features through vectorization). Dimensionality reduction techniques encompass principal component analysis (PCA, which uses linear transformations to convert high-dimensional features into linearly independent principal components, replacing the original feature set with a few key components), factor analysis (summarizing multiple correlated features into a few "factors" that represent shared trends), and autoencoders (AutoEncoders, which compress high-dimensional features into low-dimensional representations through neural networks, achieving dimensionality reduction by minimizing reconstruction errors, particularly suitable for nonlinear data).

3.4. Data reduction

Data reduction is a technique for simplifying datasets, aiming to reduce data size, improve processing efficiency, and lower costs while retaining core information and analytical value.

Dimensionality reduction techniques are divided into feature selection and feature extraction. Feature selection includes filter methods (using variance selection and other methods to screen features), wrapper methods (using models such as SVM to evaluate and select the optimal subset), and embedded selection (selecting features during model training, such as L1 regularized linear models that automatically set the coefficients of unimportant features to 0); feature extraction transforms high-dimensional features into low-dimensional ones through methods like PCA to achieve dimensionality reduction.

Data reduction methods include random sampling (suitable for scenarios with uniform data distribution), clustering reduction (suitable for scenarios with uneven distribution), regression reduction (suitable for time-series data), and histogram reduction (suitable for discrete data).

Data compression and approximate representation aim to reduce storage and transmission costs. Data compression includes lossless compression (using Huffman coding, suitable for high-precision scenarios) and lossy compression (using wavelet transform, suitable for low-precision scenarios). Methods for approximate data representation include statistical model representation, sampling representation, and summary representation (suitable for scenarios where understanding overall characteristics is needed).

4. Application analysis of data preprocessing in different industries

4.1. Financial industry: risk control and customer profile construction

Data preprocessing serves as the core infrastructure for risk control and customer profiling in the financial sector, enabling the processing of multi-source heterogeneous data [2].

4.1.1. Application of preprocessing in risk control

In risk control, a bank's personal credit risk assessment employs a preprocessing workflow: data cleaning (imputing missing values and detecting suspicious transactions), data integration (consolidating data into a unified risk dataset), data transformation (constructing standardized risk control features), and data reduction (eliminating low-relevance features while retaining key ones). The preprocessed dataset is fed into the XGBoost risk control model, which achieves an 18% improvement in bad debt prediction accuracy, effectively mitigating credit risks.

4.1.2. Application of preprocessing in customer profile construction

A securities firm's high-net-worth client profiling workflow includes three key steps: data cleansing (smoothing and anomaly detection), data integration (user data matching), and data transformation (discretization and profiling feature construction). The resulting high-net-worth client profiles enabled the firm to deliver personalized product recommendations, achieving a 22% increase in product conversion rates.

4.2. Healthcare: electronic medical record integration and disease prediction

The core challenge in data preprocessing for healthcare is the multi-source heterogeneity of electronic medical records (EMRs), with the goal of integrating data into structured "patient health records" to support disease prediction and precision medicine [2].

4.2.1. Preprocessing for electronic medical record integration

A tertiary hospital implemented the following workflow: data cleaning (imputation of missing values for "family medical history" and identification of erroneous records for "lab test indicators exceeding medical ranges"), data integration (entity recognition to match medical records across different departments and standardization of "allergy history" information), and data transformation (text vectorization to convert "medical record descriptions" and standardization of "lab test indicator" units). After integration, the system achieved "cross-department patient information sharing," resulting in a 30% improvement in physician diagnostic efficiency.

4.2.2. Disease prediction preprocessing

A medical AI company's diabetes complication prediction workflow includes data cleaning (imputing missing values for glycated hemoglobin and smoothing blood glucose fluctuation data), data integration (consolidating data such as blood glucose records), and data transformation (constructing features like the number of blood glucose exceedances in the past 3 months and removing redundant features). After preprocessing, the dataset is fed into a deep learning model, achieving an 89% accuracy rate in diabetic nephropathy prediction and a 78% accuracy rate in early 6-month warning.

4.3. E-commerce: supporting user behavior analysis and personalized recommendation

4.3.1. User behavior analysis preprocessing

An e-commerce platform's "User Repurchase Behavior Analysis" process includes three key steps: data cleaning (eliminating accidental clicks and refining the "User Reviews" field), data integration (combining behavioral data), and data transformation (creating features like "repurchase frequency over the past 30 days" and segmenting "repurchase intervals"). The analysis helps the platform identify specific user groups, enabling targeted campaigns that boosted repurchase rates by 15% [3].

4.3.2. Recommendation system preprocessing

A short video platform's workflow includes data cleaning (eliminating invalid data from "traffic brushing"), data integration (combining behaviors like "viewing"), data transformation (generating features such as "video viewing duration ratio" and eliminating dimensionality issues), and data reduction (converting to "user preference vectors"). When applied to a collaborative filtering recommendation model, this approach increased average video viewing time by 20% and user retention by 12% [4].

4.4. Intelligent manufacturing: supporting equipment fault early warning and product quality control

4.4.1. Equipment monitoring preprocessing

A car manufacturer's "welding robot fault early warning" system operates through three phases: data cleaning (smoothing vibration data and detecting abnormal temperature spikes), data transformation (constructing vibration frequency fluctuations and eliminating dimensional inconsistencies), and data reduction (removing humidity data while retaining key features). After preprocessing, the dataset is fed into an LSTM time series prediction model, achieving a 92% fault early warning accuracy and reducing downtime by 40% [5].

4.4.2. Quality control preprocessing

The "chip production quality control" process of a certain electronics manufacturing enterprise includes data cleaning (identifying unqualified samples), data integration (integrating "equipment parameter data", etc.), and data transformation (constructing features such as "equipment parameter fluctuation score", and discretizing "test data"). After preprocessing, the dataset is input into the random forest quality prediction model, which reduces the chip defective rate by 25% and improves production efficiency by 18% [6].

5. Challenges and discussions in data preprocessing

First, data quality and consistency. The "multi-source heterogeneity" of enterprise data makes issues of data quality and consistency prominent, including problems such as inconsistent formats, semantic inconsistencies, pseudo-missing data, and delays in data updates. These issues increase the cost of format conversion, lead to semantic conflicts, raise the complexity of handling missing values, and affect the timeliness of analysis results.

Second, processing efficiency and scalability with large-scale data. With the explosive growth of data volume, traditional single-machine preprocessing tools can no longer meet the needs, making distributed preprocessing an inevitable choice. However, this brings new challenges, including distributed data sharding, distributed task scheduling, real-time preprocessing requirements, and toolchain compatibility, which affect processing efficiency and scalability.

Third, privacy protection and data security compliance. Data preprocessing involves users' private data, and regulations have strict requirements for privacy protection. There are issues such as difficulties in data desensitization, privacy problems in cross-organizational preprocessing, the need for compliance audits in preprocessing processes, and high implementation costs of privacy-enhancing technologies, which affect privacy protection and compliance.

Fourth, difficulties in preprocessing multimodal and unstructured data. Multimodal and unstructured data account for more than 80% of the total enterprise data and continue to increase, making their preprocessing a new challenge. Problems include the heterogeneity of multimodal data, difficulties in the semantic understanding of unstructured data, challenges in fusion preprocessing of multimodal data, and limitations of preprocessing tools, which increase the complexity of preprocessing.

Combining current technical pain points and industry needs, the future development of data preprocessing will show trends such as automation and low-code development, AI-enhanced intelligent preprocessing, edge preprocessing and real-time processing, privacy-enhanced preprocessing technologies, unification of multimodal preprocessing frameworks, and end-to-end integration of preprocessing and analysis.

6. Conclusion

This paper systematically reviews the key contents of data preprocessing in big data analysis, comprehensively covering its fundamental importance, mainstream technical methods, industry practical applications, and many current challenges. The study points out that data preprocessing is a core foundational link to ensure the accuracy and reliability of subsequent data analysis results, and occupies a crucial position in the entire data value chain. Among them, key technologies such as data cleaning, data integration, data transformation, and data reduction cooperate with and support each other, jointly constructing a systematic preprocessing process. They effectively solve problems such as noise, redundancy, inconsistency, and missing data in raw data, and significantly improve the overall quality and availability of data. Through in-depth analysis of multiple typical industry scenarios such as financial risk control, medical and health care, and e-commerce, this paper reveals the key role and actual effects of data preprocessing technology in real business environments, demonstrating its important value in improving decision-making accuracy and business efficiency. However, with the continuous growth of data volume, increasingly complex structures, and the increasing proportion of unstructured data, the current data preprocessing process still faces huge challenges in terms of processing efficiency, data security and privacy protection, and the level of automation and intelligence. There are obvious bottlenecks, especially when dealing with high-dimensional, streaming, and cross-domain data. It should be noted that this paper is mainly based on the induction and review of existing literature, and has not introduced a detailed comparative analysis of specific empirical cases. This limitation may affect the practical guiding significance and promotion applicability of the conclusions. Looking forward to the future, relevant research should pay more attention to developing automated preprocessing frameworks with strong adaptability and controllable costs, and under the premise of fully ensuring data privacy and compliance, improve the comprehensive processing capability of multimodal and heterogeneous data, thereby promoting the in-depth implementation and value release of big data analysis in a wider range of industries.

References

[1] China Academy of Information and Communications Technology. (2024). Big Data White Paper 2024. Beijing: China Academy of Information and Communications Technology.

[2] Coursera. (2024). Big Data Analysis Specialization (University of California, San Diego). Retrieved from https: //www.coursera.org/specializations/big-data-analysis.

[3] Provost, F., & Fawcett, T. (2022). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking (2nd Edition). Sebastopol: O'Reilly Media.

[4] Prakash, A. S., & Pullela, P. K. (2024). Understanding Big Data Analysis: Framework for Exploring Traits, Business Values, and Challenges in E-commerce. AIP Conference Proceedings, 3121(1), 040025. https: //doi.org/10.1063/5.0221484.

[5] Gou, X. L. (2019). Fault Diagnosis Method Analysis of Power Transmission and Distribution Equipment Based on Big Data Mining Technology. Telecom Power Technology, 36(1), 282-282.

[6] Gartner. (2024). SASE Will Improve Your Distributed Security Everywhere. Stamford: Gartner.