Applications and Challenges of Artificial Intelligence in Data Cleaning

Pusen Gao
The University of Melbourne, VIC 3000, Melbourne, Australia

Abstract: Data quality issues have long been a key factor affecting model performance and the reliability of decision-making. As a critical component of data preprocessing, data cleansing is gradually shifting from traditional rule-based and statistical methods toward AI-driven automated approaches. This paper provides a systematic review of the application of artificial intelligence in data cleansing. It classifies and analyzes relevant methods across three dimensions-machine learning, deep learning, and large language models-and introduces an evaluation method based on data quality dimensions. The paper compares and analyzes different technologies in terms of completeness, accuracy, and consistency, and summarizes their application effectiveness in tasks such as anomaly detection, missing value imputation, and entity matching.

Keywords: Data Cleaning; Artificial Intelligence; Machine Learning; Deep Learning; Large Language Models; Data Quality; Missing Value Imputation; Anomaly Detection

1. Introduction

Against the backdrop of rapid advancements in big data and artificial intelligence, data quality has become a critical factor influencing data analysis and model performance. However, real-world data often contains issues such as missing values, noise, outliers, and inconsistencies, which severely compromise the accuracy of decision-making. Consequently, data cleansing, as a crucial step in the data processing workflow, has garnered widespread attention.

Traditional data cleaning methods primarily rely on statistical analysis and rule-based techniques, which perform well when handling small-scale or structured data. However, when dealing with large-scale and complex data, these methods suffer from low efficiency and poor scalability. With the advancement of machine learning and deep learning, researchers have proposed various automated data cleaning methods, such as generative model-based missing value imputation and representation learning-based outlier detection, significantly enhancing the intelligence of data cleaning. In recent years, the rise of large language models (LLMs) has further propelled advancements in this field. Their strengths in semantic understanding and cross-task processing have unlocked new potential for complex data cleaning tasks.

Data cleaning refers to the process of detecting and correcting errors, incompleteness, or inconsistencies in raw data, and it is a crucial component of data preprocessing. Based on the type of data issues, data cleaning tasks can generally be categorized into three types: outlier detection, missing value imputation, and entity resolution.

Outlier detection aims to identify data points that deviate from normal patterns. Common methods include density-based algorithms such as the Local Outlier Factor (LOF), which compares the density of a data point with that of its neighbors [1]. Missing value imputation focuses on completing incomplete data, while entity matching addresses identifying the same entity across different data sources.

To evaluate the effectiveness of data cleansing more systematically, researchers typically assess results from the perspective of data quality. Common data quality metrics include completeness, accuracy, and consistency. Among these, completeness reflects the presence of missing data, accuracy measures the correctness of data values, and consistency focuses on whether there are conflicts or contradictions between different data sets. These metrics not only provide quantitative standards for evaluating data cleansing effectiveness but also serve as a foundation for comparing different technical approaches.

2. AI-Based Data Cleaning Methods

2.1 Machine Learning Methods

Traditional machine learning methods are widely used in data cleaning tasks, particularly in anomaly detection and missing value imputation. For example, the LOF algorithm identifies outliers based on local density differences [1]. These methods are effective for structured data and offer relatively low computational cost.

2.2 Deep Learning Methods

Deep learning models can capture complex patterns and improve the automation and accuracy of data cleaning. In missing value imputation, generative models such as variational autoencoders and GANs can learn latent data distributions and produce high-quality imputations. Studies have shown that deep learning models often outperform traditional methods, especially in large-scale and high-dimensional datasets [2,3].

In anomaly detection tasks, autoencoder-based models identify anomalous samples by analyzing reconstruction errors, effectively capturing complex patterns. Furthermore, deep learning methods have made progress in entity matching tasks; for instance, by learning features from both text and structure, they achieve more accurate entity alignment.

2.3 LLM-based Methods

Large Language Models (LLMs) offer a new technical approach to data cleaning. Compared to traditional machine learning and deep learning methods, LLMs possess stronger semantic understanding and cross-task generalization capabilities, enabling them to handle more complex data quality issues with ease, including semantic inconsistencies, data transformation, and the integration of multi-source data.

For example, recent studies have demonstrated that LLMs can automatically detect and correct data errors by combining semantic reasoning with statistical inference [4]. In domain-specific scenarios such as healthcare, LLMs can incorporate contextual knowledge to improve cleaning accuracy [5].

Compared to traditional methods, LLMs possess distinct advantages in handling unstructured data and semantically related issues. For instance, they can reasonably impute missing values based on contextual reasoning or identify semantic conflicts in data through natural language understanding. Furthermore, LLMs support more automated data cleaning workflows, reducing reliance on manually designed rules and thereby improving data processing efficiency.

3. Evaluation of Different AI-driven Data Cleaning

3.1 Data Quality Dimension

To systematically evaluate the effectiveness of artificial intelligence (AI) technologies in data cleansing, this paper begins by analyzing data quality dimensions. Common data quality metrics include completeness, accuracy, and consistency. Specifically, completeness addresses data missingness, accuracy measures how closely data values align with actual values, and consistency reflects whether conflicts or contradictions exist among data points.

Different AI methods exhibit varying performance across various data quality metrics. For example, basic machine learning methods demonstrate high stability in anomaly detection for structured data but have limited performance on complex semantic problems; deep learning methods, by learning data distributions, have advantages in improving data accuracy; while large language models demonstrate stronger capabilities in semantic consistency and complex data repair.

These dimensions are widely used in prior research. For example, *Statistical Analysis with Missing Data* emphasizes the importance of completeness and accuracy in reliable analysis [6]. Similarly, systematic reviews highlight accuracy and consistency as key evaluation metrics in data cleaning research [7].

3.2 Step-by-Step Analysis Framework

During the data cleansing process, artificial intelligence technologies are typically applied at various stages. To more systematically describe their mechanisms of action, this paper divides the data cleansing process into three stages: problem detection, problem repair, and quality evaluation.

In the detection stage, machine learning and deep learning methods identify anomalies. In the repair stage, generative models and LLMs are used to correct errors. In the evaluation stage, results are assessed using data quality metrics.

This staged framework aligns with existing systems such as HoloClean, which integrates error detection and repair [8], and DataXFormer,

which automates transformation and cleaning processes [9].

3.3 Comparative Analysis of Methods

Based on the aforementioned data quality dimensions and step-by-step framework, a comprehensive comparison of different AI methods can be conducted. Overall, machine learning methods offer the advantages of simple implementation and low computational costs, but their capabilities are limited in complex data scenarios; deep learning methods excel in modeling high-dimensional data and capturing nonlinear relationships, but they are highly dependent on data scale and computational resources; large language models, meanwhile, possess significant advantages in semantic understanding and cross-task processing, enabling more automated data cleaning workflows.

However, in practical applications, different methods often need to be used in combination. For example, with structured data, machine learning methods can be employed for preliminary anomaly detection, followed by the use of deep learning or large language models for data repair and optimization. Therefore, building a data cleansing framework that integrates multiple methods has become a key direction for improving overall data quality.

4. Challenges and Limitations

Although artificial intelligence (AI) technology has demonstrated significant advantages in data cleansing, it still faces multiple challenges in practical applications.

First, the issue of explainability remains a major factor limiting the application of AI methods. In particular, deep learning and large language models often lack transparency in their decision-making processes, making it difficult for users to understand why the models make specific modifications to the data. In high-risk sectors such as healthcare and finance, this "black box" nature may undermine system credibility, thereby limiting their practical deployment.

Second, AI methods are highly dependent on the quality of the data itself. If bias or errors exist in the training data, models may learn inaccurate patterns, thereby amplifying existing issues during the data cleansing process. This problem is particularly pronounced in deep learning and large language models, as their performance relies heavily on large-scale training datasets.

Furthermore, the generalization capabilities of models remain limited. Many methods perform well on specific datasets but experience a significant drop in performance when applied across domains or under different data distributions. For example, in fields such as healthcare or finance, where data structures and semantic complexity are high, existing models often struggle to transfer their knowledge directly, leading to inconsistent cleaning results.

Large language models also face unique challenges. Their outputs are inherently uncertain and may produce data correction results that do not align with reality (known as the "hallucination" phenomenon), thereby compromising data quality. Furthermore, these models typically require substantial computational resources, posing challenges in terms of cost and efficiency.

Finally, privacy and security concerns cannot be overlooked. When processing data containing sensitive information, the data cleaning process may involve data sharing and model training, thereby posing risks of privacy breaches. Therefore, how to ensure data security while improving data quality has become a key direction for future research.

5. Future Directions

With the continuous advancement of artificial intelligence technology, there remains vast scope for research in the field of data cleansing. First, the deep integration of large language models with data cleansing processes may emerge as a key direction; for example, the introduction of "human-in-the-loop" mechanisms is expected to enhance the reliability and controllability of results while maintaining a high degree of automation.

Furthermore, the development of Explainable AI (XAI) is crucial for enhancing the transparency of data cleansing systems, particularly in highly sensitive sectors such as healthcare and finance, where it helps build user trust. At the same time, privacy protection technologies will emerge as a key research focus to ensure data security while improving data quality. Finally, improving the generalization capabilities of models across diverse data scenarios to enable cross-domain data cleansing applications remains a critical challenge to be addressed in the future.

6. Conclusions

This paper presents a systematic review of the application of artificial intelligence in data cleansing. It classifies and analyzes relevant methods across three dimensions-machine learning, deep learning, and large language models-and compares their effectiveness based on data quality metrics. The study demonstrates that AI technologies can significantly enhance the automation and accuracy of data cleansing in tasks such as anomaly detection, missing value imputation, and semantic repair, with deep learning and large language models exhibiting superior performance in complex data scenarios.

However, existing methods still face certain limitations in terms of interpretability, generalization, and stability; in particular, the issue of uncertainty associated with large language models in practical applications requires further investigation. Therefore, future work should focus on enhancing system reliability and security while improving model performance.

Overall, artificial intelligence offers new approaches and technical pathways for data cleansing and holds significant potential for improving data quality. However, further methodological refinements and system optimizations are needed to advance its practical implementation and development.

References

[1] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD.

[2] Wang, Z., Akande, O., Poulos, J., & Li, F. (2021). Are deep learning models superior for missing data imputation in large surveys? arXiv.

[3] Camino, R. D., Hammerschmidt, C. A., & State, R. (2019). Improving missing data imputation with deep generative models. arXiv.

[4] Zhu, Y., et al. (2024). Large language models for data cleaning. arXiv.

[5] Lee, J., et al. (2025). Leveraging large language models for clinical data cleaning. arXiv.

[6] Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). Wiley.

[7] Côté, P.-O., et al. (2024). Data cleaning and machine learning: A systematic literature review. arXiv.

[8] Rekatsinas, T., et al. (2017). HoloClean: Holistic data repairs with probabilistic inference. VLDB.

[9] Hildebrandt, T., et al. (2016). DataXFormer: Data cleaning and transformation for data integration. IEEE ICDE.

