Dean&Francis
ISSN 2959-6157

Application of Random Forest Technique in Data Cleaning

Wenhao Bai

School of Statistics, University of International Business and Economics, Beijing, China
Corresponding Author: 202492001@uibe.edu.cn

Abstract:

Data cleaning is an important part of data preprocessing. Its role is to ensure the accuracy of data analysis and improve model performance. Its main tasks are to identify and resolve various problems in raw data, such as missing values, outliers, and duplicate records, thereby providing strong support for subsequent data mining and modeling. In the era of big data, traditional data cleaning methods suffer from poor adaptability and low accuracy when dealing with high-dimensional, nonlinear, and massive datasets. By contrast, random forest has been widely applied in data cleaning because of its resistance to overfitting, strong ability to handle high-dimensional data, low requirement for complex preprocessing, and relatively strong interpretability. This paper reviews recent domestic and international studies on the principles, methods, and research status of random forest in the three core scenarios of data cleaning. It mainly discusses its advantages over traditional methods, its existing limitations, and future research directions. At the same time, visualized content is incorporated to improve the readability of this review and to provide references for future research and applications.

Keywords: Random forest; Data cleaning; Missing values

1. Introduction

As technologies such as big data and artificial intelligence continue to evolve rapidly, data has become the primary production factor for the development of various industries. "Data-driven" has gradually become a mainstream model for industrial transformation and business innovation. However, during the collection, transmission, and storage of raw data, various complex factors such as insufficient equipment accuracy, human operational errors, diverse and heterogeneous data sources, and system noise can affect the data, inevitably resulting in some "dirty data", including missing data, abnormal data, and duplicate data. According to industry statistics, the proportion of dirty data in real data sets can reach 30% to 50%. If not directly used for modeling and analysis, it will significantly affect the correctness of the results, leading to decision-making deviations and even strategic errors [1,2].

Therefore, data cleaning is a key part of data preprocessing. By using various technical means to iden-

tify and correct errors, remove duplicate data, and fill in missing values, the accuracy and completeness of the data are significantly enhanced, laying a solid foundation for high-quality data analysis. The traditional data cleaning methods mainly use statistical strategies (mean interpolation, median interpolation, 3σ criterion for identifying outliers) and manual rule methods (format verification, range check) [3]. Although the principles are simple and the operation cost is low, they also have obvious drawbacks, namely, it is difficult to discover complex nonlinear relationships and high-dimensional interaction effects between variables, cannot well adapt to large-scale and high-dimensional data, and is highly dependent on expert experience and the definition of artificial rules, with low processing efficiency and poor scalability [4,5].

Random Forest is a representative algorithm proposed by machine learning scholar Breiman in 2001 based on decision tree integration. It uses bootstrap sampling to establish multiple decision trees and adopts voting or averaging to obtain the final result, having good anti-overfitting ability and high prediction robustness. This algorithm does not require complex feature engineering and data preprocessing, and can provide variable importance assessment, improving the interpretability of the model. Therefore, Random Forest has been widely applied in data cleaning tasks, such as filling in missing values, anomaly detection, and noise identification, providing a new methodological support for processing complex, high-dimensional, and nonlinear data [1,2].

This paper conducts a systematic review using literature research and comparative analysis methods. Through systematic retrieval of high-quality academic literature in relevant fields at home and abroad, common application scenarios of Random Forest in data cleaning processes, main technological development situations, and current encountered problems are sorted out, and a systematic comparison is made with traditional data cleaning methods in terms of effect, efficiency, robustness, and interpretability, etc., to establish a systematic review framework. This paper not only systematically organizes and summarizes the research results of Random Forest in data preprocessing but also provides certain theoretical support and practical references for industrial sectors to carry out data quality management, having high academic value and practical significance.

The innovations of this review can be summarized as follows. Firstly, this paper constructs a systematic review framework for the application of random forest in data cleaning by integrating three core tasks, namely missing value imputation, outlier detection, and duplicate data identification, into one unified analytical perspective. Secondly, it provides a comparative analysis between traditional data cleaning methods, basic random forest methods, and improved random forest methods from multiple dimensions, including accuracy, efficiency, adaptability, interpretability, and implementation convenience. Lastly, this paper combines literature review with visualized process and performance comparison, which helps present the application logic, practical advantages, current limitations, and future development directions of random forest in data cleaning more clearly.

2. Problems in Data Cleaning and Limitations of Traditional Methods

The key point of data cleaning lies in dealing with three situations: missing values, outliers, and duplicate data. The traditional approach has obvious shortcomings, mainly including the following points.

Missing values are the most common problem in data quality, caused by collection failures, transmission losses, or human negligence, which will result in a decrease in sample size and incomplete data, thereby affecting the reliability of the analysis. The traditional processing methods include deletion and interpolation. The deletion method is simple to operate, but it discards useful information. The interpolation method (such as mean, linear interpolation, etc.) only considers the statistical characteristics of a single attribute and does not take into account the correlation between data, resulting in low interpolation accuracy and poor generalization ability [3].

Outliers are caused by equipment errors, misoperations, or actual abnormal events, which will affect the statistical characteristics and thereby affect the accuracy of the model. They are divided into pseudo-outliers that need to be deleted and true outliers that need to be retained. The traditional detection methods include statistical methods (3σ criterion, box plot) and distance methods (KNN). The former is only suitable for single-dimensional data with normal distribution and is prone to misjudgment; the latter is computationally complex and cannot be used for a large amount of data, and it is difficult to distinguish between pseudo-outliers and true outliers [4].

Duplicate data is divided into complete duplication and semantic duplication. Due to repeated collection, entry errors, etc., duplicate records are caused, resulting in increased storage costs and decreased processing efficiency. The traditional identification methods include rule matching method and hash matching method. The rule matching method requires manual writing of rules, is inefficient, and has poor universality. The latter is only suitable for completely duplicated data and has poor recognition effect for semantic duplication, and is prone to misjudgment due

ISSN 2959-6157

to hash conflicts [5].

3. Applications of Random Forest in Data Cleaning

To present the overall logic of data cleaning based on random forests more clearly, its workflow can be summarized as a closed-loop process, as shown in Fig. 1. Firstly, the raw data undergoes preprocessing and quality diagnosis. Then, three major data quality issues - missing values, outliers, and duplicate records - are identified and handled separately. Subsequently, a quality assessment step is conducted on the cleaned data to determine whether the results meet the required standards. If the results are satisfactory, these data can be used for subsequent analysis and modeling; otherwise, the process will return to the previous stage for further revision. Because of its own advantages, random forest has been widely applied in the three major scenarios of missing value imputation, outlier detection, and duplicate data identification. Many researchers have further improved the algorithm to enhance both cleaning efficiency and accuracy [2]. Visualized content can clearly demonstrate its application logic and performance.

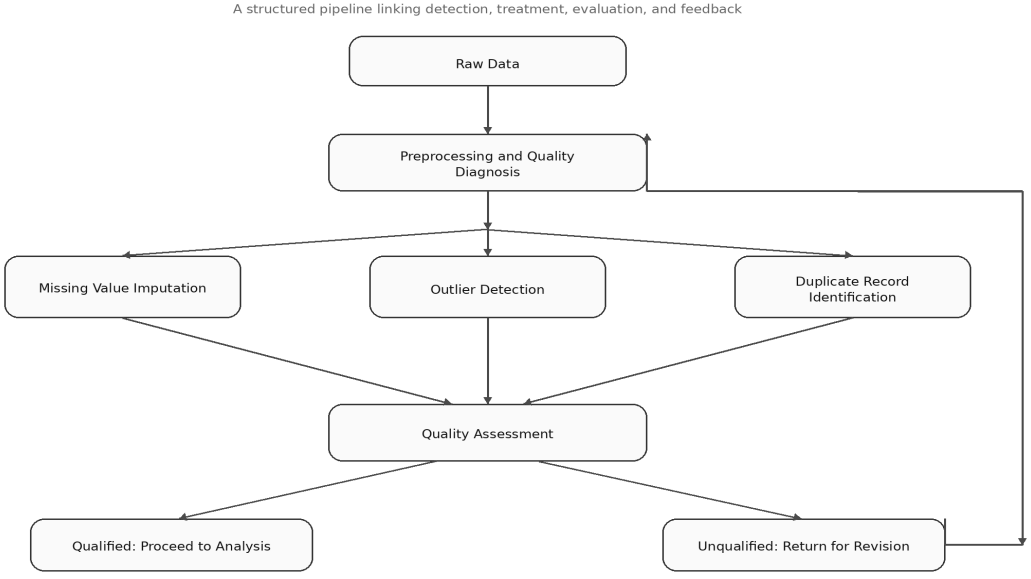

Fig. 1 Workflow of Random Forest Data Cleaning

The core principle of missing value imputation is to treat the attribute containing missing values as the target variable and use the other attributes as features to build a random forest regression or classification model for predicting missing values. This approach can capture complex nonlinear relationships in the data and reduce errors through ensemble prediction. The main steps include data preprocessing, model training, missing value prediction, and iterative optimization. Feature importance is also used to select important features so as to improve imputation accuracy. In related studies, Han Honggui and others proposed an improved random-forest-based imputation method for wastewater treatment data, and its imputation error was much lower than that of traditional random forest imputation, reduced by 15% to 25% [3]. Related improvements involving temporal features and spatial correlations have also been reported in time-series and environmental-health datasets [2,6].

The core principle of outlier detection and treatment is to identify outliers based on the distributional differences between abnormal samples and normal samples. There are both supervised and unsupervised modes. The main steps include feature extraction, model training, anomaly detection, and classification. Studies have shown that random-forest-based hybrid methods can achieve high detection accuracy and lower false alarm rates in practical anomaly-detection tasks [4,7].

In duplicate data identification and removal, the main idea is to transform the identification problem into a classification problem, construct similarity feature vectors, and use a random forest model to determine whether records are duplicates. This approach can reflect semantic relationships and does not require manually defined rules. The main process includes feature construction, sample labeling, model training, identification, and deduplication. Existing studies indicate that combining random forest with edge computing, hash matching, or deep feature extraction can improve both recognition efficiency and semantic duplicate identification accuracy [5,8].

From the perspective of visualization, the random-for-

est-based data cleaning process forms a closed loop. Raw data first undergoes preprocessing and quality diagnosis, then random forest is applied separately to the three kinds of problems, and finally quality assessment is used to determine whether the results are qualified. Qualified results are used for subsequent analysis, while unqualified ones require reprocessing. Comparative results across different methods show that random forest has clear advantages over traditional algorithms in terms of imputation error and detection accuracy, and improved random forest methods perform even better.

4. Comparison Between Random Forest and Traditional Methods

To more intuitively compare the differences among traditional data cleaning methods, basic random forests, and methods based on improved random forests, this study evaluated their performance from five dimensions: accuracy, speed, adaptability, interpretability, and implementation convenience, as shown in Fig. 2. Overall, the methods based on random forests demonstrated significant advantages in accuracy and adaptability, especially when dealing with high-dimensional, non-linear, and complex datasets [2]. The improved random forests further enhanced processing efficiency while maintaining high cleaning performance. In contrast, traditional methods still had certain advantages in implementation convenience, but their applicability in complex data cleaning tasks was relatively limited.

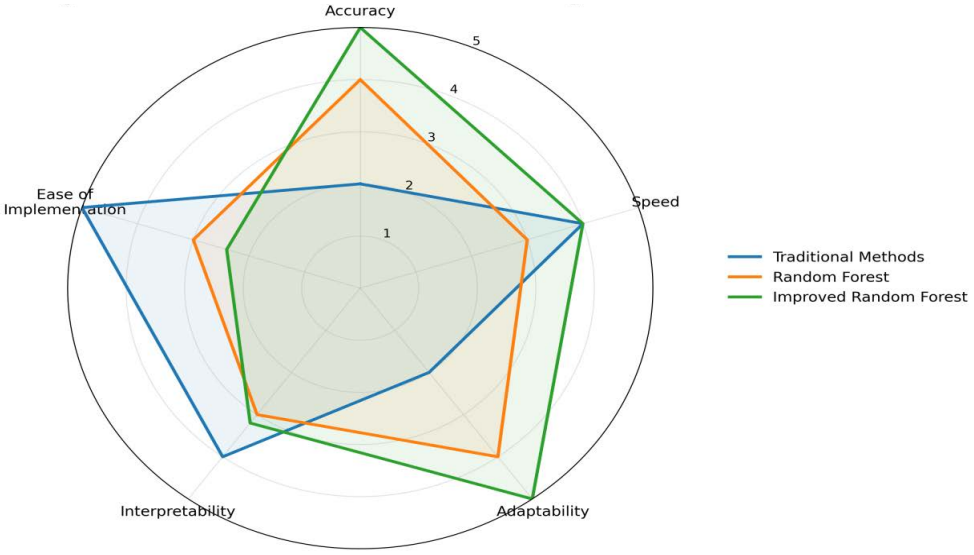

Fig. 2 Performance Comparison of Data Cleaning Methods

In terms of processing accuracy, random forest is much more accurate than traditional methods. It performs significantly better in handling missing value imputation, with the error controlled within 15%, and reduced to below 10% after improvement. Traditional methods, by contrast, often exceed this range, for example reaching about 25%. In outlier detection, random forest can achieve an accuracy above 90%, and after improvement the accuracy exceeds 95%. In duplicate data recognition, the accuracy is above 94%, and rises to above 97% after improvement [3,4,8].

In terms of processing efficiency, traditional methods perform much better than basic random forest, but worse than improved random forest. The traditional method has lower computational cost and is suitable for small samples and low-dimensional data. The basic random forest, however, is slow when dealing with large-scale high-dimensional data because it constructs multiple decision trees. Nevertheless, through methods such as parallel computing and edge computing, the improved algorithm can meet the requirements by increasing processing speed (by more than 40%) [2,5].

In terms of adaptability, random forest is stronger than traditional algorithms. It can be directly applied to high-dimensional, non-linear data without complex preprocessing and can adapt to fields such as environment, industry and Internet of Things. It can effectively handle various types of missing values and outliers, while the application scope of traditional methods is relatively narrow [2,6].

From the perspective of interpretability, random forest is not as easy to explain as traditional rule-based methods, but is superior to other machine learning methods such as neural networks. It can explain the cleaning results through feature importance, while traditional rule-based

ISSN 2959-6157

methods are more transparent during the process but have poorer generalization [8].

In terms of ease of implementation, traditional methods are simple in principle and easy to code, making them suitable for non-specialists. Random forest, by contrast, requires knowledge of machine learning and parameter tuning, making it more difficult to implement. However, with the development of relevant frameworks, this difficulty has gradually decreased.

Overall, random forest is suitable for large-scale, complex, and massive datasets requiring advanced cleaning, while traditional methods are more suitable for simple data. The actual choice should depend on the characteristics of the dataset and the business requirements.

5. Research Limitations and Future Development Trends

At present, the application of random forest in data cleaning still has several limitations. First, its efficiency is not high enough; the large computational load in massive-data and real-time environments is not favorable for applications in edge computing and the Internet of Things. Second, there are no effective methods for handling non-random missing values, and the imputation performance remains unsatisfactory due to the lack of targeted approaches. Third, its adaptability to multi-source heterogeneous data is insufficient, and feature fusion is not yet effective enough. Fourth, its ability for automatic parameter tuning is weak and still mainly relies on human experience, which leads to low efficiency. Fifth, its interpretability remains insufficient, making it difficult to clearly explain why particular cleaning results are produced, which limits its applicability in fields such as healthcare and finance [9].

According to current technological trends, future development may mainly proceed in the following directions. First, processing efficiency should be improved through parallel computing, edge computing, and lightweight model design so as to meet the requirements of real-time cleaning. Second, methods for handling non-random missing values should be improved by incorporating domain knowledge into sampling and feature selection. Third, the matching capability for multi-source heterogeneous data should be strengthened by exploring feature fusion and transfer learning. Fourth, parameter self-adaptive optimization should be improved by introducing reinforcement learning, Bayesian optimization, and related methods to reduce labor costs. Fifth, interpretability should be enhanced by using visualization to present the decision process and by constructing evaluation systems. Sixth, the scope of multi-scenario integrated applications should be expanded so as to build an integrated preprocessing system and extend it to new fields [10].

6. Conclusion

Because of its strong resistance to overfitting, strong adaptability, and good interpretability, random forest effectively overcomes the limitations of traditional data cleaning methods. It performs well in the three scenarios of missing value imputation, outlier detection, and duplicate data identification. Its imputation accuracy and detection accuracy are much higher than those of traditional cleaning methods, making it suitable for various complex data cleaning scenarios in fields such as environment, industry, and the Internet of Things. Current research still has problems such as insufficient efficiency and limited applicability. At the same time, this review shows that the practical value of random forest lies not only in improving cleaning accuracy, but also in providing a more flexible solution for complex and heterogeneous data environments. In the future, with algorithmic improvement and technological integration, development will move toward greater efficiency, precision, self-adaptation, interpretability, and multi-scenario integration, thereby providing strong support for high-quality data processing in the era of big data and enabling truly data-driven decision-making. Therefore, the future development of this field should place equal emphasis on model performance, computational efficiency, and interpretability.

References

[1] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32.

[2] Xiao Yang, Wang Xinzhang, Peng Cheng, Chen Junfeng, Jiang Tao. Data cleaning method for operational data of offshore oil and gas production equipment based on improved random forest. Modern Chemical Research, 2023, (12): 155-157.

[3] Han Honggui, Zhao Zifan, Wu Xiaolong, Yang Shiheng, He Zheng, Zhao Nan. Data cleaning method for urban wastewater treatment process operation data based on improved random forest. Journal of Beijing University of Technology, 2021, 47(05): 421-430.

[4] Wei Tai, He Shaoxiong, Hu Ziwu, Cao Lixin. Cleaning of abnormal data from wind turbines based on an improved isolation forest algorithm. Science Technology and Engineering, 2024, 24(09): 3691-3699.

[5] Cao Ying. Research on industrial sensing data cleaning and compression method based on elastic edge computing. Wuhan University of Technology, 2022.

[6] Liu Yue, Hao Shuxin, Liu Jie, Xu Dongqun. Research

on a data cleaning framework for environmental health risk assessment data. Journal of Environmental Hygiene, 2025, 15(10): 878-884+907.
[7] Hou Dengyun, Nan Xinyuan, Li Hailong. Data cleaning method for wastewater treatment process based on adaptive DBSCAN-LOF. Journal of Northeast Normal University (Natural Science Edition), 2025, 57(03): 47-55.
[8] Wang Jing. Research on pollutant discharge data cleaning method for catalytic cracking units based on isolation forest and neural network. China University of Petroleum (Beijing), 2020.
[9] Chen Hongqiao. Big data cleaning algorithm for the Internet of Things based on isolation forest. Information Recording Materials, 2025, 26(01): 154-156+165.
[10] Sun Ruizao, Wei Lu. Cleaning of abnormal wind power data based on isolation forest and standard deviation detection. Henan Science, 2023, 41(03): 313-320.