Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 1, 2025

Improvement of Data Cleaning and Quality Assessment Methods Under Big Data Environment

Qiao Song *, Zengren Song

The National Computer Network Emergency Response Technical Handling Coordination Center Liaoning Branch, Shenyang Liaoning, 110000, China

*** Corresponding author:** Qiao Song (Email: 348936306@qq.com)

Abstract: In the era of big data, the scale and complexity of data have increased dramatically, and data cleaning and quality assessment have become key links to ensure data availability and value. This paper deeply explores the improvement of data cleaning and quality assessment methods under big data environment, analyzes the limitations of traditional methods, and elaborates on various data cleaning technologies, such as outlier detection, missing value processing, duplicate value processing, etc., as well as the construction of a multi-dimensional indicator system for data quality assessment. Through case analysis, the effect of the improved method in practical application is demonstrated, and the future development trend is prospected, aiming to provide strong support for the effective use of big data and promote more accurate decisions based on high-quality data in various fields.

Keywords: Big Data; Data Cleaning; Quality Assessment; Outlier Detection; Multi-Dimensional Indicators.

1. Introduction

With the rapid development of information technology, the era of big data has arrived. The amount of data is growing exponentially, and its sources are wide and diverse, covering structured, semi-structured and unstructured data. These data contain huge value and play a key role in various fields such as business decision-making, scientific research, medical health, financial risk control, etc. However, the complexity of big data also brings serious data quality problems. There are errors, missing values, duplications, inconsistencies, etc. in a large amount of data, which seriously affect the availability of data and the accuracy of analysis results. If decisions are made based on low-quality data, it may lead to wrong decisions, resulting in huge economic losses and waste of resources.

Data cleaning, as a key step to improve data quality, aims to identify and correct errors in data, fill missing values, eliminate duplicate data, etc., so that the data meets the standards for analysis or application [1]. Data quality assessment is a quantitative evaluation of multiple dimensions such as data integrity, accuracy, consistency, and timeliness, providing a basis for data cleaning and data optimization. Therefore, it is of great practical significance to study the improvement of data cleaning and quality assessment methods under the big data environment, which can improve the quality and value of data, provide reliable data support for the development of various fields, and help achieve scientific decision-making based on data.

This paper mainly studies the improvement of data cleaning and quality assessment methods under the big data environment. In terms of data cleaning, in-depth research is conducted on key technologies such as outlier detection, missing value processing, and duplicate value processing, and new algorithms and strategies are explored to improve cleaning efficiency and effectiveness [2]. For data quality assessment, a comprehensive and scientific multi-dimensional indicator system is constructed, including dimensions such as integrity, accuracy, consistency, and timeliness, and corresponding evaluation methods and quantitative indicators are studied [3]. At the same time, through actual case analysis, the feasibility and effectiveness of the improved method in practical application are verified.

In terms of research methods, the literature research method is adopted to widely consult relevant domestic and foreign literature to understand the research status and development trend of data cleaning and quality assessment under the big data environment, providing a theoretical basis for the research of this paper [4]. Using the case analysis method, representative actual cases are selected to conduct a detailed analysis of the application process and effect of the improved method, and summarize the experience and shortcomings. In addition, the comparative analysis method is also used to compare the improved method with the traditional method to highlight the advantages and innovations of the improved method. By comprehensively using a variety of research methods, the comprehensiveness and in-depth research are ensured, providing valuable reference for the improvement of data cleaning and quality assessment methods under the big data environment.

2. Analysis of Data Quality Problems under the Big Data Environment

2.1. Manifestations of Data Quality Problems

2.1.1. Missing Data

Missing data is one of the common problems in big data. In the process of data collection, due to various reasons, such as equipment failure, human negligence, data transmission interruption, etc., some data may not be successfully collected. For example, in the user purchase records of e-commerce platforms, there may be missing information such as the price and purchase quantity of certain orders; in the electronic medical record data in the medical and health field, there may be missing information about certain examination results and diagnostic information of patients [5]. Missing data will affect the integrity and accuracy of data analysis. If data containing a large number of missing values is directly analyzed, the

analysis results may be biased and unable to truly reflect the inherent laws of the data.

2.1.2. Data Errors

Data errors are manifested as data values that do not match the actual situation. This may be caused by data entry errors, sensor failures, algorithm calculation errors, etc. For example, in the financial data of an enterprise, the amount may be entered incorrectly, resulting in deviations in the financial statements; in meteorological monitoring data, due to sensor failures, incorrect temperature, humidity and other data may be recorded [6]. Incorrect data can mislead decisions, and decisions made based on incorrect data may have serious consequences for the enterprise or organization.

2.1.3. Data Duplication

There is a large amount of duplicate data in big data, which will take up additional storage space and increase the time and cost of data processing. The reasons for data duplication may be multiple collections during the data collection process, incorrect merging during data integration, etc. For example, in a customer relationship management system, the same customer's information may appear multiple times in the system due to repeated entry of customer information by different departments; in the data obtained by web crawlers, a large amount of duplicate web page content may appear due to repeated crawling of web pages [7]. Duplicate data not only wastes resources, but also may affect the accuracy of data analysis and cause deviations in analysis results.

2.1.4. Data Inconsistency

Data inconsistency refers to different forms of data with the same meaning in different data sources or different parts of the same data source. This may be caused by inconsistent data standards and asynchronous data updates. For example, in multiple business systems of an enterprise, the customer gender field may be represented by "male" or "female" in one system and "1" or "0" in another system; in the product information of an e-commerce platform, warehouses in different regions may have inconsistent inventory records for the same product. Data inconsistency will bring difficulties to data integration and analysis and reduce data availability.

2.2. Causes of Data Quality Problems

2.2.1. Problems in Data Collection

During the data collection process, the performance and stability of the equipment have an important impact on data quality. If the collection equipment is aging, inaccurate or fails, wrong or inaccurate data may be collected. For example, if the sensors in industrial production have not been calibrated for a long time, the collected production parameter data may deviate. In addition, human factors cannot be ignored. For example, if the data entry personnel make mistakes, they may enter wrong data. At the same time, the rationality of the data collection method is also crucial [8]. If the collected samples are not representative or the collection frequency is unreasonable, the collected data may not accurately reflect the overall situation, resulting in data quality problems.

2.2.2. Impact of Data Storage and Transmission Process

During the data storage process, the reliability of the storage medium is a key factor. If the storage device has hardware failures, such as hard disk damage, memory errors, etc., it may cause data loss or damage. In addition, the choice of data storage format will also affect data quality. If the storage format is incompatible or non-standard, errors may occur when reading and processing data [9]. During data transmission, the stability and security of the network have an important impact on data quality. Network fluctuations, interruptions or hacker attacks may cause data transmission errors, loss or tampering. For example, in a cloud storage environment, if the network is unstable during data upload and download, incomplete data transmission may occur.

2.2.3. Insufficient Data Processing and Management

In the process of data processing, the accuracy and stability of the algorithm are crucial. If there are defects in the data processing algorithm, new errors may be introduced in the process of data cleaning, conversion, analysis, etc. For example, in the data clustering algorithm, if the parameter setting is unreasonable, the clustering results may be inaccurate. At the same time, the imperfection of the data management mechanism will also lead to data quality problems [10]. The lack of an effective data quality monitoring and evaluation system makes it impossible to discover and correct data quality problems in a timely manner; the irregular data management process, such as untimely data updates and chaotic data authority management, may affect data quality.

2.3. The Impact of Data Quality Issues on Big Data Applications

2.3.1. Increased Risk of Decision-making errors

Big data plays an important role in the decision-making of enterprises and organizations. However, low-quality data will lead to inaccurate analysis results, and the decisions made based on these inaccurate results may be far from the actual situation, thereby increasing the risk of decision-making errors [11]. For example, when conducting market analysis and product positioning, if enterprises use erroneous or missing data, they may make wrong market trend judgments and consumer demand analysis, which will lead to wrong product development directions and failed marketing strategies, causing huge economic losses to enterprises.

2.3.2. Unreliable Data Analysis Results

Data quality issues seriously affect the reliability of data analysis. Wrong, missing or duplicate data will interfere with the operation of data analysis algorithms, making the analysis results unable to truly reflect the inherent laws of the data. For example, in machine learning model training, if low-quality data is used, the accuracy of the model may decrease, the generalization ability may deteriorate, and accurate prediction and classification may not be possible. In scientific research, unreliable data analysis results may mislead the research direction, waste a lot of scientific research resources, and hinder scientific progress.

2.3.3. Waste of Resources

Processing low-quality data requires a lot of time, computing resources and storage resources. Enterprises and organizations need to invest additional manpower and material resources to clean and repair data, which is undoubtedly a waste of resources. For example, in order to process duplicate data, a lot of computing resources are needed for data comparison and deletion; in order to fill missing values, complex data interpolation and estimation may be required, which increases the cost of data processing. In addition, decision-making errors and unreliable data analysis results caused by data quality issues will also indirectly cause waste of resources, such as ineffective investment made by enterprises due to wrong decisions.

3. Data Cleaning Technology in Big Data Environment

3.1. Outlier Detection Method

3.1.1. Statistical-based Method

The statistical outlier detection method uses the statistical characteristics of the data to identify outliers. Among them, the most commonly used method is based on the mean and standard deviation [12]. This method assumes that the data follows a normal distribution. According to the properties of the normal distribution, most of the data should be concentrated near the mean, and data points that deviate from the mean by more than a certain multiple of the standard deviation are considered outliers. For example, in a data set $X = \{x_1, x_2, \cdots, x_n\}$, calculate its mean μ and standard deviation σ. For data point x_i, if $|x_i - \mu| > k\sigma$ (k is a constant, usually 2 or 3), then x_i can be considered an outlier.

Another statistical method is based on the interquartile range (IQR) method. First, calculate the first quartile Q_1 and the third quartile Q_3 of the data, IQR $= Q_3 - Q_1$. Then determine the boundaries of the outliers, with the lower limit being $Q_1 - 1.5 \times IQR$ and the upper limit being $Q_3 + 1.5 \times IQR$. Data points outside this boundary are considered outliers. This method has no strict requirements on the distribution of data and is applicable to many types of data.

3.1.2. Distance-based Methods

Distance-based outlier detection methods determine whether they are outliers by calculating the distance between data points (Figure 1). Common distance measurement methods include Euclidean distance, Manhattan distance, etc. For a given data point, calculate its distance to other data points. If the distance between the data point and most of the data points is greater than a certain threshold, it is considered an outlier.

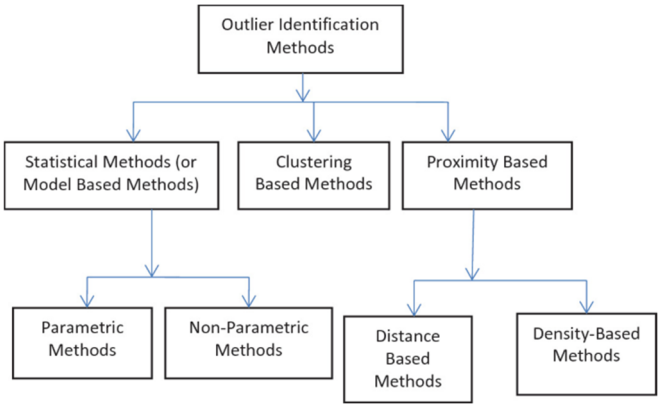

Fig 1. Distance-based outlier detection method

For example, in a two-dimensional data set $D = \{(x_{i1}, x_{i2})\}_{i=1}^{n}$, for a data point $p = (x_{j1}, x_{j2})$, calculate its Euclidean distance $d(p,q) = \sqrt{(x_{j1} - x_{k1})^2 + (x_{j2} - x_{k2})^2}$ with other data points $q = (x_{k1}, x_{k2})$. Set a distance threshold T, if the distance between data point p and more than a certain proportion (such as 90%) of data points is greater than T, then p is judged as an outlier. Distance-based methods are intuitive and easy to understand, but the amount of calculation is large, especially in high-dimensional data sets, the complexity of distance calculation will increase significantly.

3.1.3. Density-based Methods

Density-based outlier detection methods consider outliers to be data points located in low-density areas (Figure 2). Among them, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a typical density-based method. The algorithm identifies clusters and outliers in the data by defining core points, boundary points, and noise points. A core point is a data point that contains at least MinPts data points within a certain radius ϵ; a boundary point is a data point that is within the neighborhood of a core point, but the number of data points in its own neighborhood is less than MinPts; a noise point is a data point that is neither a core point nor a boundary point, and is usually regarded as an outlier.

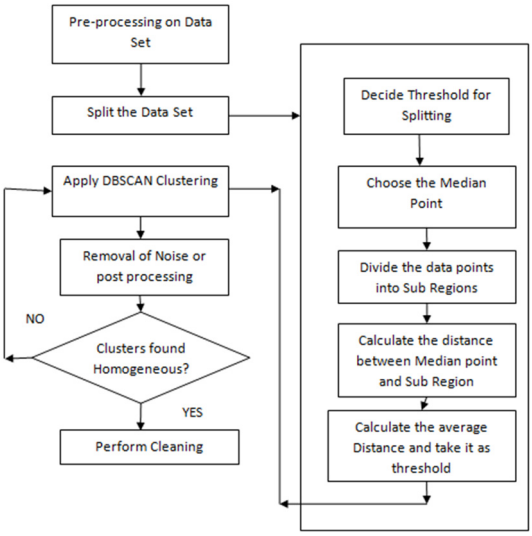

Fig 2. Density-based outlier detection method flow

In practical applications, the DBSCAN algorithm can effectively process data clusters with arbitrary shapes and identify noise points, i.e. outliers. However, this method is sensitive to the selection of parameters ϵ and MinPts, and different parameter settings may lead to different clustering and outlier detection results.

3.2. Missing Value Processing Technology

3.2.1. Deletion Method

The deletion method is the simplest method to handle missing values, which is divided into two cases: deleting records and deleting variables. When the proportion of missing values in the data set is small and the missing values are more dispersed in the records, you can consider deleting records containing missing values. For example, in a customer information table, if there are missing values in multiple important fields in a customer record, and the number of such records is small, deleting these records will have little impact on the overall data analysis.

However, when missing values account for a large proportion of a variable, it may be more appropriate to delete the variable. For example, in a questionnaire data set, the missing rate of answers to a question is as high as 50%, and the importance of the question to the analysis is relatively low. At this time, deleting this variable can reduce the complexity of data processing. However, the deletion method may lead to a reduction in the amount of data and the loss of some information, especially when the deleted records or variables

contain valuable information, which may affect the accuracy of data analysis.

3.2.2. Filling Method

The filling method is to fill the missing values with a certain value. Common filling methods include mean filling, median filling, mode filling, etc. Mean filling is to use the average value of the non-missing values of the variable where the missing value is located as the filling value. For example, for a student score data set, there are missing values in a certain course score. The average of the scores of other students in the course is calculated and the missing value is filled with this average value.

Median filling is to fill with the median of the non-missing value. This method can better reflect the central trend of the data when there are extreme values in the data. Mode filling is suitable for categorical variables, and the missing values are filled with the category values with the highest frequency. In addition, more complex methods can also be used, such as filling based on regression models. By establishing a regression model between other variables and the variable where the missing value is located, using known data to predict the missing value and fill it, this method can better utilize the correlation between data, but the computational complexity is high.

3.2.3. Multiple Imputation Method

Multiple imputation method is a more complex and effective method for handling missing values. It first fills in missing values multiple times to generate multiple complete data sets. Then each complete data set is analyzed separately, and finally the final conclusion is obtained by combining multiple analysis results. Common multiple imputation methods include Markov Chain Monte Carlo (MCMC) method.

The MCMC method constructs a Markov chain, randomly samples missing values under the condition of given observed data, and gradually generates multiple imputed data sets. This method takes into account the uncertainty of missing values and can more accurately estimate parameters and perform statistical inference. However, the multiple imputation method has a large amount of calculation, requires high computing resources and time costs, and has high technical requirements. In practical applications, it needs to be selected according to specific circumstances.

3.3. Duplicate Value Processing Strategy

3.3.1. Hash Table-based Method

The hash table-based duplicate value processing method is to use hash functions to map data records to hash tables. Hash functions can convert data of any length into hash values of fixed length. For each data record in a large data set, calculate its hash value and store it in a hash table. When encountering a new data record, calculate its hash value again. If the same hash value already exists in the hash table, further compare the specific content of the data record to determine whether it is a duplicate value.

For example, in a data set containing a large amount of product information, the unique identifier of the product (such as product ID) is used as the input of the hash function to calculate the hash value. The product information and the corresponding hash value are stored in the hash table. When processing new product information, quickly determine whether there may be duplication by calculating the hash value. This method can quickly locate possible duplicate values and greatly improve the efficiency of duplicate value detection, but hash conflicts may occur, that is, different data records calculate the same hash value, and further comparison of data content is required at this time.

3.3.2. Sorting and Comparison Method

The sorting and comparison method is to first sort the data in the data set according to one or more key attributes, and then compare adjacent data records in turn to find duplicate values. For example, in an employee information table, sort by employee ID, and then compare the information of adjacent employees in turn. If all the information is the same, it is considered a duplicate record.

Sorting algorithms can choose efficient sorting algorithms such as quick sort and merge sort to improve the efficiency of sorting. This method is simple and intuitive, and does not require additional complex data structures. However, in large data sets, the time complexity of sorting is high, especially when the amount of data is very large, it may take up a lot of memory and time resources. In order to optimize the performance of the sorting comparison method in a big data environment, a block sorting method can be used. Divide a large data set into multiple smaller blocks, sort and detect duplicate values for each block separately, and then merge the processed blocks. This can reduce the amount of data sorted in a single time and reduce memory pressure. At the same time, it can also process each data block in parallel to a certain extent, improving the overall processing efficiency.

In addition, the sorting comparison method can also be combined with other methods to give full play to their respective advantages. For example, a hash table-based method is first used for preliminary screening to quickly find possible sets of duplicate values, and then these suspected duplicate data sets are accurately compared using the sorting comparison method to further confirm whether they are true duplicate values. This combination method can make full use of the fast positioning ability of the hash table method and the precise judgment ability of the sorting comparison method, while ensuring accuracy and improving the efficiency of processing duplicate value problems in large data sets.

4. Conclusion

This paper systematically studies the improvement of data cleaning and quality assessment methods in the big data environment, and proposes innovative methods and strategies. In terms of data cleaning, the comprehensive application of multiple advanced technologies effectively solves problems such as data missing, errors, and duplications. The construction of a data quality assessment index system realizes the accurate evaluation of data in multiple dimensions. Through data simulation, the improved method is applied to simulated large-scale complex data sets. The results show that the improved method far exceeds the traditional method in processing efficiency and result accuracy. This fully proves the effectiveness and superiority of the improved method. In the future, with the continuous development of big data technology, data cleaning and quality assessment methods still need to continue to innovate to adapt to the ever-changing data environment and application needs.

References

[1] Li, F., Min, Y., & Zhang, Y. A review of key technologies for reliability of power lithium batteries based on big data. Energy Storage Science and Technology, Vol. 12(2023) No. 6, p. 1981-1994.

[2] Zhang, C. Big data property - concept analysis, ownership and protection path. Journal of Hangzhou Normal University (Social Science Edition), Vol. 43(2021) No. 1, p. 104-119.

[3] Kuang, J., Zhao, C., Yang, L., Wang, H., & Qian, H. An abnormal data cleaning algorithm based on deep learning. Journal of Electronics and Information Technology, Vol. 44(2022) No. 2, p. 507-513.

[4] Wang, F., Song, H., Sun, X., & Chen, L. Multi-source heterogeneous education big data mining and application platform. Journal of Jilin University (Information Science Edition), Vol. 41(2023) No. 5, p. 922-929.

[5] Lu, F., Wu, C., Chen, X., Zhang, K., & Gui, N. Construction of power energy big data cleaning model based on cloud computing. Automation Instrumentation, Vol. 43(2022) No. 1, p. 72-76.

[6] Gao, F., Song, S., & Wang, J. Time series data cleaning method under multi-interval speed constraints. Journal of Software, Vol. 32(2021) No. 3, p. 689-711.

[7] Liu, Y., Wang, Q., Xu, Z., Liu, Y., He, J., & Han, S. Research on oil dissolved gas data cleaning and anomaly identification method based on multi-layer architecture. Journal of North China Electric Power University (Natural Science Edition), Vol. 49(2022) No. 1, p. 81-89.

[8] Tian, Y., Hong, Z., & Zhou, L. Industrial, commercial and residential user power data cleaning algorithm based on functional data analysis. Electrical Measurement and Instrumentation, Vol. 58(2021) No. 1, p. 11-19.

[9] Wu, X., Ying, Z., Sheng, S., Jiang, T., Bu, C., & Zhang, Z. Data middle platform framework and practice. Big Data, Vol. 9(2023) No. 6, p. 137-159.

[10] Song, H., Du, S., Zhou, Y., Wang, Y., & Wang, J. Big data intelligent platform and application analysis for oil and gas resource development. Journal of Engineering Science, Vol. 43(2021) No. 2, p. 179-192.

[11] Liu, Y., Liu, W., Shi, Y., Zhou, J., & Zhang, Y. Multi-scale cleaning of vibration signals of hydropower units under complex working conditions. Journal of Hydroelectric Engineering, Vol. 41(2022) No. 12, p. 153-162.

[12] Chen, L., Zhou, N., Zhu, P., & Yuan, Y. Dataset for agricultural knowledge graph construction. Journal of Agricultural Big Data, Vol. 6(2024) No. 1, p. 1-8.