Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 12, No. 3, 2024

Research on Data Quality Analysis Based on Data Mining

Jieting Lian

New York University, New York, US

Abstract: As the process of digitalization continues to advance, big data plays a pivotal role across various industries. However, the quality of data directly impacts the results of data analysis and the accuracy of decision-making. High-quality data can provide a reliable basis for decisions, whereas poor-quality data may lead to erroneous judgments and cause incalculable losses. Therefore, the management and enhancement of data quality have become central issues in contemporary data management. Data mining technology, as a powerful tool, excels not only in extracting valuable information from large-scale datasets but also in demonstrating significant potential in improving data quality. By mining and analyzing the patterns and characteristics hidden within data, it is possible to identify and rectify defects, thereby enhancing the overall quality of data and increasing the reliability of data-driven decision-making. This paper delves into the definition and dimensions of data quality, explores the fundamental principles of data mining technology, and examines its specific applications in data quality analysis, aiming to offer new insights and methods for data quality management.

Keywords: Data mining; data quality; analysis.

1. Introduction

In today's information-driven society, data has become a crucial asset for organizations. However, issues with data quality are pervasive, affecting the efficacy of data-driven decision-making. Problems such as data omission, inconsistencies, and duplicate records can lead to skewed analytical outcomes and erroneous decisions. While traditional data quality management methods can alleviate these issues to some extent, they often fall short when dealing with large-scale, multidimensional, and dynamically evolving datasets. Data mining techniques have increasingly emerged as a vital tool for data quality analysis and management due to their ability to automatically uncover latent patterns and anomalies within vast amounts of data. These techniques not only effectively identify errors and irregularities in data but also enhance data structure through predictive models and clustering algorithms, thereby improving consistency and accuracy. Exploring the application of data mining technology in data quality management is not merely an enhancement of existing methods but represents a significant innovative direction in the field of data management, offering substantial practical value [1].

2. Definition and Dimensions of Data Quality

Data quality is a fundamental pillar in the process of data analysis and decision-making, directly affecting the credibility and efficacy of data applications. Its core dimensions encompass accuracy, completeness, consistency, timeliness, and reliability. Accuracy pertains to the ability of data to faithfully reflect objective reality, representing a fundamental requirement of data quality; completeness emphasizes the comprehensiveness of data, where missing information may lead to asymmetries that impact the reliability of analysis results; consistency ensures that data remains coherent across different sources and time points, avoiding conflicts and misinterpretations; timeliness requires data to reflect the most current status, which is crucial in real-time decision-making scenarios; reliability refers to the stability of data throughout its acquisition, storage, and transmission processes, free from external disruptions. These dimensions are interrelated and collectively determine the overall quality of data. In the era of big data, with the diversification and complexity of data sources, the demand for high standards of data quality has become increasingly apparent, necessitating the development of systematic evaluation and enhancement methods. This creates extensive opportunities for the application of data mining techniques in data quality management.

3. Overview of Data Mining Techniques

3.1. Fundamentals of Data Mining

The essence of data mining lies in the automated extraction of valuable patterns, rules, or knowledge from vast quantities of data. This process is grounded in statistics, machine learning, and database systems, uncovering potential correlations and regularities through exploration and analysis. The crux of data mining is its ability to reveal latent knowledge within the data, which may manifest in forms such as classification, clustering, association rules, or anomaly detection. Classification techniques are extensively employed in supervised learning, wherein models are built to predict the categories of unknown data based on the analysis of data with known categories. Clustering, a method used in unsupervised learning, organizes data into subsets with strong homogeneity, proving particularly crucial in customer segmentation and market analysis. Association rule mining aims to uncover intriguing relationships between data items, with shopping basket analysis being a typical application. Anomaly detection identifies rare events that deviate from the majority of data, holding significant importance in fraud detection and cybersecurity. The power of data mining lies in its ability to extract insights from complex and voluminous data that are difficult for humans to discern directly, enabling

organizations to make more precise and efficient decisions. However, the efficacy of data mining hinges on the quality of data and the preliminary data preprocessing work; poor-quality data can lead to model bias and misleading conclusions [2]. Thus, while data mining enhances data value, it also faces severe challenges and complexities, necessitating ongoing research and refinement.

3.2. Classification Algorithms Major Data Mining Techniques and Methods

3.2.1. Classification Algorithms

Classification algorithms are among the most prevalent techniques in data mining, primarily employed to address supervised learning tasks with the aim of categorizing data into predefined classes based on its characteristics. A widely favored method is the decision tree, esteemed for its intuitive nature and interpretability. Decision trees learn rules from data through a tree-like structure, ultimately generating a model that can predict the categories of new data. However, decision trees are prone to overfitting, which can result in poor performance when encountering new data. To mitigate this issue, the random forest approach was developed. Random forests enhance classification accuracy and stability by constructing multiple decision trees and aggregating their results. Its robust generalization capability proves particularly effective for complex, high-dimensional data. Nonetheless, random forests exhibit higher computational complexity and their results are less intuitively interpretable compared to a single decision tree. On the other hand, Support Vector Machines (SVM) represent a classification technique based on maximizing the margin between classes. By identifying an optimal hyperplane, SVMs can segregate data points into distinct categories. For data that is not linearly separable, SVMs utilize kernel functions to map data into a higher-dimensional space, achieving more precise classification. While SVMs excel in scenarios with small samples and high dimensionality, their efficiency may decrease when dealing with large-scale data. The choice of classification algorithm should be aligned with the characteristics of the data and the specific requirements of the task. Each algorithm possesses its own strengths and limitations, and the key lies in selecting the appropriate method based on the complexity of the problem and the nature of the data to achieve accurate and efficient classification.

3.2.2. Clustering algorithm

Clustering techniques play an indispensable role in data mining, particularly in the realm of unsupervised learning. Their purpose is to partition data objects into multiple clusters with similar characteristics, thereby unveiling the intrinsic structure within the data. Among the most commonly used clustering algorithms is K-means, which optimizes through iteration by assigning data points to the nearest cluster centers and continuously updating these centers' positions until convergence is achieved. The advantage of K-means lies in its high computational efficiency, capacity to handle large datasets, and relatively straightforward implementation. However, K-means is sensitive to the initial selection of centroids and can easily become trapped in local optima. Furthermore, K-means assumes that clusters are spherical and of similar size, which may be inadequate for handling clusters of complex shapes or varying sizes. In contrast, hierarchical clustering constructs a dendrogram to progressively cluster data, without the need to predefine the number of clusters. Hierarchical clustering can be categorized into agglomerative and divisive types, where the former starts from individual data points and merges them progressively, while the latter begins with the whole dataset and divides it incrementally. The strength of hierarchical clustering lies in its ability to produce a comprehensive hierarchical structure, making it well-suited for scenarios requiring a deep understanding of data stratification [3]. However, hierarchical clustering's computational complexity is relatively high, resulting in lower efficiency when handling large-scale data, and once a step is completed, it cannot be reverted or adjusted. Both K-means and hierarchical clustering exhibit unique value in different contexts. K-means is ideal for processing large-scale, relatively simple datasets, whereas hierarchical clustering excels in scenarios demanding a profound understanding of data structure. Choosing the appropriate clustering algorithm necessitates a thorough consideration of the data's nature and the analysis objectives to effectively reveal hidden patterns and structures within the data.

3.2.3. Association Rules

Association rule mining is a pivotal technique in data mining, aimed at uncovering implicit associations between itemsets within vast datasets. This technique finds extensive application in market basket analysis, assisting businesses in optimizing product placement and promotional strategies by revealing purchasing relationships between various products. The Apriori algorithm stands as one of the most classical methods in association rule mining, effectively discovering potential associations through the iterative expansion of frequent itemsets. However, the efficiency of the Apriori algorithm is impacted by the process of generating frequent itemsets, with its computational complexity increasing exponentially, particularly when handling large-scale datasets. To enhance the efficiency of association rule mining, the FP-growth algorithm was developed. FP-growth compresses the dataset by constructing a frequent pattern tree (FP-tree), thereby avoiding the cumbersome candidate itemset generation process found in Apriori. FP-growth demonstrates superior efficiency and scalability when dealing with large-scale data. Nonetheless, the complexity associated with the construction and manipulation of FP-trees necessitates substantial memory usage, and FP-growth may encounter performance bottlenecks in high-dimensional or sparse datasets. When selecting an association rule mining algorithm, it is crucial to balance efficiency and complexity while considering the specific characteristics of the dataset [4]. The Apriori algorithm performs well in scenarios with relatively few rules, whereas FP-growth is more suited for large-scale, complex datasets. Regardless of the chosen method, the essence of association rule mining lies in revealing latent relationships within the data to drive data-driven decisions. However, this process demands higher data quality, with accuracy and completeness being prerequisites for mining effective association rules. Consequently, the continuous improvement of algorithms to address the challenges posed by diverse datasets remains a significant research focus in the field of data mining.

3.2.4. Abnormal detection

Anomaly detection plays a pivotal role in data mining, particularly when identifying rare or unusual patterns within datasets. The Isolation Forest, a method grounded in the principles of random forests, employs a strategy of progressively isolating data points by randomly selecting features and split values. Because anomalous points are more easily separated, the Isolation Forest is adept at efficiently

identifying outliers. It boasts high computational efficiency, making it suitable for large-scale datasets, and imposes minimal assumptions about data distribution. Nonetheless, when dealing with high-dimensional data, the Isolation Forest may encounter challenges related to performance, necessitating careful parameter adjustment to balance detection accuracy with computational complexity. On the other hand, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) represents another prevalent anomaly detection method. It utilizes density clustering to group points in high-density regions into clusters while designating points in low-density regions as anomalies. DBSCAN's strengths lie in its ability to identify clusters of arbitrary shape and its effectiveness in detecting noise points without requiring a predefined number of clusters [5]. However, DBSCAN is sensitive to parameter selection, particularly in scenarios with significant variations in data density, making it difficult to determine suitable parameter values. Additionally, its performance may be constrained when handling high-dimensional data. In practical applications of anomaly detection, choosing the appropriate algorithm is crucial. The Isolation Forest is well-suited for large-scale, feature-rich environments, whereas DBSCAN excels in scenarios requiring the identification of complex cluster structures. It is important to note that anomaly detection demands not only robust algorithms but also a profound understanding of the data. Factors such as noise, incompleteness, and distribution characteristics can significantly impact the effectiveness of anomaly detection, rendering it both a tool and a vital measure of data quality in the analysis of data quality.

4. Application of Data Mining in Data Quality Analysis

4.1. Data Preprocessing and Cleaning

In practical data processing, raw data often suffers from noise, missing values, duplicate records, and inconsistencies, which can significantly impact subsequent analysis and modeling. Consequently, data preprocessing is not only a means to enhance data quality but also a fundamental prerequisite for ensuring the efficacy of data mining. The core task of data cleaning is to identify and rectify errors and anomalies within the data. For missing values, common approaches include removing records with missing data, imputing missing values using means or medians, or employing more sophisticated interpolation methods to estimate the missing values. This approach must be tailored to the specific dataset and application scenario; excessive deletion may lead to loss of information, while simplistic imputation methods might obscure underlying patterns in the data [6]. Thus, a balance must be struck between maintaining data integrity and optimizing model performance. Handling noise is also a critical task in data cleaning, particularly in contexts such as sensor data or user input data, where noise often manifests as outliers. To effectively remove noise, common techniques include smoothing filters and regression analysis, complemented by anomaly detection methods such as isolation forests or density clustering, which can identify and eliminate or correct extreme data points. During noise removal, excessive cleaning might result in the loss of valuable outlier information, especially in certain analytical scenarios where outliers themselves hold significant meaning. Therefore, noise handling requires a deep understanding of the data and must be balanced against specific application objectives. Duplicate records and inconsistencies are additional significant factors affecting data quality. The presence of duplicate records can skew analysis results toward specific data points, leading to bias. Common strategies for handling duplicates include deduplication methods based on primary keys or specific attributes. However, inconsistencies are more complex, often involving data source integration and standardization of different formats, typically necessitating a combination of domain knowledge and technical solutions. Data preprocessing is not merely a process of correcting and optimizing data but also a means of deeply understanding and evaluating data quality. Through data cleaning, potential issues can be uncovered and appropriate measures taken to improve data quality. Data preprocessing and cleaning are not only preliminary steps in data mining but are integral throughout the entire data analysis process, continually influencing the accuracy and reliability of data analysis. Only through effective data preprocessing can data mining techniques reach their full potential, truly providing valuable support for decision-making [7].

4.2. Detection of data quality problems

Data quality issues typically manifest as missing values, noise, duplicate data, outliers, and inconsistencies. These challenges may arise from errors during the data collection process, oversights during data entry, or conflicts encountered during data integration. The timely identification and rectification of data quality problems are prerequisites for ensuring data reliability and the accuracy of analyses. In addressing these issues, data mining techniques offer a multitude of effective methodologies for detection. Anomaly detection techniques, such as Isolation Forest and Density Clustering, can discern elements in a dataset that deviate from normative patterns; these anomalies often serve as indicators of potential data errors or exceptional events. Furthermore, statistical analysis methods play a crucial role in assessing data quality, as they allow for the identification of biases and anomalies through the examination of data distribution, thereby facilitating the detection of underlying issues. For instance, employing distribution analysis and box plots can swiftly pinpoint outliers and extreme values, leading to further quality scrutiny. Data consistency checks represent another vital domain, particularly prominent in the integration of multi-source data. Utilizing pattern matching and rule-checking techniques effectively identifies formatting errors and inconsistencies within the data. For example, in the realm of Geographic Information Systems (GIS), discrepancies in coordinate systems may hinder accurate data integration. Consistency checks enable the timely identification of such issues, allowing for appropriate transformations or corrections to be implemented. In practical applications, the detection of data quality problems relies not merely on isolated techniques but rather necessitates a comprehensive analysis incorporating various methods. A combination of automated detection tools and manual verification can effectively identify issues within large-scale datasets and offer remediation suggestions [8]. Nonetheless, data quality challenges often exhibit complexity and obscurity, particularly within high-dimensional and unstructured data. This demands that data analysts possess not only technical proficiency but also a profound understanding and keen insight into the data. Through effective detection of data

quality issues, one can enhance the accuracy and completeness of data, thereby laying a robust foundation for subsequent analyses. This process is an integral component of data mining efforts, with its outcomes directly influencing the scientific validity and effectiveness of data-driven decision-making. Thus, the detection of data quality problems transcends a mere technical task; it constitutes a pivotal element within the realms of data governance and analytical processes, warranting ongoing attention and in-depth exploration.

4.3. Data quality assessment and improvement

Data quality assessment and improvement are indispensable elements in the field of data mining, directly influencing the reliability and scientific validity of data-driven decision-making. The primary task of data quality assessment is to determine the performance of data in terms of accuracy, completeness, consistency, timeliness, and usability. Through assessment, one can systematically identify deficiencies within the data and provide clear directions for its enhancement. The process of evaluating data quality typically relies on a multidimensional metric system. For instance, accuracy can be assessed by comparing data against known standards or external trusted data sources, while completeness can be measured by examining the extent of missing attributes within the dataset. Consistency issues primarily involve the validation of logical relationships between data, particularly in the context of integrating multiple data sources, where ensuring matching across different systems or tables is crucial. Additionally, timeliness and usability are significant dimensions of assessment, with the former reflecting the frequency of data updates and validity periods, and the latter focusing on whether data can be efficiently accessed and utilized by users. Following the data quality assessment, addressing identified issues through improvement is an essential step. Data cleansing serves as a fundamental approach to enhancing data quality, involving the application of appropriate strategies to address issues such as missing values, noise, and duplicate data [9]. During this process, improvement measures should not only consider the characteristics of the data itself but also account for business needs to prevent the loss of critical information during cleansing. Data quality improvement is not a one-time task but a dynamic process requiring ongoing monitoring and optimization. As data application evolves through different stages, quality requirements may change, necessitating regular reassessment and adjustment of improvement strategies. Furthermore, leveraging automation tools and machine learning algorithms can further enhance the efficiency and precision of data quality improvement. For example, machine learning models can predict potential issues within the data and automatically generate repair suggestions, which is especially valuable in large-scale data processing scenarios. In summary, data quality assessment and improvement are not only processes for enhancing the value of data usage but also crucial for ensuring the reliability of data-driven decisions. Only with high-quality data can data mining truly realize its potential, providing accurate and reliable analytical support for various application scenarios [10].

5. Conclusion

Data quality is the cornerstone of data-driven decision-making, and the advent of data mining technologies has ushered in a revolutionary transformation in the analysis of data quality. The application of data mining techniques allows for the more efficient detection, evaluation, and enhancement of data quality issues, providing robust support for precise decision-making across various industries within the realm of big data. Nevertheless, challenges in data quality management persist, such as issues arising from dynamic data environments and considerations of data privacy and security, which warrant further research and exploration. In the future, with the continuous advancement of artificial intelligence and big data technologies, data mining techniques will play an increasingly pivotal role in data quality management, offering more intelligent and automated solutions for achieving high-quality data management.

References

[1] Luebbers D, Grimmer U, Jarke M. Systematic development of data mining-based data quality tools[C]//Proceedings 2003 VLDB conference. Morgan Kaufmann, 2003: 548-559.

[2] Deng W, Wang G. A novel water quality data analysis framework based on time-series data mining[J]. Journal of environmental management, 2017, 196: 365-375.

[3] Nie G, Zhang L, Liu Y, et al. Decision analysis of data mining project based on Bayesian risk[J]. Expert Systems with Applications, 2009, 36(3): 4589-4594.

[4] Haghverdi A, Öztürk H S, Cornelis W M. Revisiting the pseudo continuous pedotransfer function concept: Impact of data quality and data mining method[J]. Geoderma, 2014, 226: 31-38.

[5] Sheng V S, Provost F, Ipeirotis P G. Get another label? improving data quality and data mining using multiple, noisy labelers[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008: 614-622.

[6] Batini C, Cappiello C, Francalanci C, et al. Methodologies for data quality assessment and improvement[J]. ACM computing surveys (CSUR), 2009, 41(3): 1-52.

[7] Peña-Ayala A. Educational data mining: A survey and a data mining-based analysis of recent works[J]. Expert systems with applications, 2014, 41(4): 1432-1462.

[8] Jeihouni M, Toomanian A, Mansourian A. Decision tree-based data mining and rule induction for identifying high quality groundwater zones to water supply management: a novel hybrid use of data mining and GIS[J]. Water Resources Management, 2020, 34: 139-154.

[9] DeRosa M. Data mining and data analysis for counterterrorism[M]. Washington, DC: CSIS Press, 2004:11.

[10] Berti-Equille L. Measuring and modelling data quality for quality-awareness in data mining[M]//Quality measures in data mining. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 101-126.