World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

The impact of data preprocessing on data mining outcomes

Bhargavi Konda *

Systems Analysts, HRIS, Atrius Health, USA.

World Journal of Advanced Research and Reviews, 2022, 15(03), 540-544

Publication history: Received on 16 August 2022; revised on 18 September 2022; accepted on 20 September 2022

Article DOI: https://doi.org/10.30574/wjarr.2022.15.3.0931

Abstract

Data preprocessing is a vital initial step during knowledge discovery because it determines the success of data mining projects. A dataset's quality and representation stand as the primary element because any presence of redundant, irrelevant, too noisy, or unreliable information will severely disrupt the knowledge discovery process. The preprocessing phase first converts unstructured data into an analytical format alongside solutions for data inconsistencies, errors, and missing values to maintain data mining result integrity. The preprocessing corrects data quality problems and arranges data properly, improving data mining model accuracy, efficiency, and interpretability. The data mining pipeline requires data preprocessing as its essential foundation since it provides multiple techniques to convert raw data into an effective analytical format. Data mining depends heavily on preprocessing operations because they guarantee proper analysis results through accurate correction of errors and optimal data structure development and absent data point management.

Keywords: Data Mining; Big data; Data Preprocessing; Data Analytics; Model-based Imputation

1. Introduction

Data mining techniques are used to mine big data for valuable information. Data mining is commonly used in various industries with the growth of big data. The availability and influence of big data and their analytics have enabled new growth opportunities and created new challenges [1]. Advanced data mining processes are required to process big data. Employee benefit planning begins with essential data preprocessing steps, determining how well data mining activities proceed [2].

Test data quality and representation are crucial because any dataset problems with redundancy, irrelevancy, noisiness, or unreliability will create significant challenges in knowledge discovery [4]. Data preprocessing converts unprocessed data into analytical-ready data while resolving inconsistencies and errors and handling missing values, affecting data mining results' integrity. Adapting datasets for various data mining algorithms requires data preprocessing, which transforms difficult-to-use information into valuable resources [3].

1.1. Data Preprocessing Techniques

Data preprocessing executes multiple methods that enhance data quality and mining suitability. The preprocessing tools for medical tests involve data cleaning, transformation, reduction, and integration [4,5]. Data cleaning includes several operations addressing missing data, fixing inconsistent data points, and eliminating unwanted irregularities [5]. Different approaches exist to address missing values in data by using mean, median, or mode values for imputation and through advanced approaches such as k-nearest neighbors and model-based imputation.

* Corresponding author: Bhargavi Konda

Standardization and normalization methods adjust the data range to maintain equal influence on analysis between attributes of varying size. The analysis benefits from data reduction methods, which select features while reducing dimensions because this practice enhances computational speed and prevents overfitting errors. Data integration uses multiple source data to unify datasets through inconsistency resolution and redundancy reduction, creating one complete dataset for analysis [5].

According to Crone et al. (2005), data mining performance heavily depends on implementing correct data preprocessing methods. Selecting a domain marks the beginning of determining appropriate data collection sources. The target data should stem from the original dataset to enhance result reliability. The obtained target data needs data preprocessing for practical application. The prepared dataset becomes ready for analysis to generate results by implementing mining techniques [6]. The improvement of data mining model accuracy, efficiency, and interpretability results from effectively addressing data quality issues during suitable data formatting tasks in preprocessing.

2. Literature Review

Research demonstrates how data preprocessing is a primary component of knowledge discovery and mining operations. Many research investigations establish that when preprocessing methods are appropriately chosen for application, they enhance data mining algorithms' effectiveness across multiple application domains. A reliable and effective output in data mining depends heavily on proper data preprocessing execution.

The most suitable combination of preprocessing methods does not yield universal results because it depends uniquely on each dataset and its corresponding data mining objectives. Available preprocessing algorithms do not produce reliable and effective results with consistent performance on every dataset. Before being used in data mining models, data must undergo data cleaning operations, including missing value management and outlier detection, to achieve acceptable input quality. Most standard data mining models break down without missing values due to the inability of such values to participate in comparative operations and categorization or arithmetic computations. Data mining models need proper attention to missing values as a priority step before implementation.

Data preprocessing aims to affect multiple dimensions of data mining outcomes in ways that influence model accuracy besides enhancing efficiency and improving both interpretability and generalizability. Ongoing research devoted to data preprocessing development remains crucial for data mining progress since it permits better discovery of knowledge from data. Research into existing documents reveals that data preprocessing is essential to knowledge discovery and mining. Multiple research studies prove that applying suitable preprocessing methods substantially enhances data mining algorithm operations throughout many fields [4]. Implementing correct data preprocessing methods produces improved final data mining outputs and better reliability [6].

Data preprocessing approaches differ according to different datasets and mining tasks, so the correct combination needs assessment for every specific case [4]. Access to numerous preprocessing solutions does not resolve current difficulties in reaching stable, reliable performance across datasets [4]. Data processing requires techniques that handle missing values and detect outliers because they ensure the suitability of data as an input for data mining models. Most data mining models require complete datasets since unreported values need to be usable in assessment functions or cannot qualify for classification but also need to serve in calculations. Skilled data analysts should handle missing values before implementing data mining models because it makes a critical difference [7,8,9].

3. Methodology

An experimental framework enabled researchers to evaluate data mining result changes based on the implementation of data preprocessing methods. The framework delivers an extensive approach to obtain and cleanse data before conducting a data mining examination. The research utilized a multi-source dataset acquisition process, which included financial, healthcare, and social media data, to achieve complex data and diverse characteristics. The chosen dataset selection criteria included data sets with different sizes, diverse dimensions, and varying noise levels to guarantee the general applicability of research results.

The data preprocessing followed a uniform sequence of techniques that combined data cleaning with data transformation, data reduction, and data integration functions. The assessment of data preprocessing impacts on different analytical tasks employed a set of data mining algorithms that covered classification regression and clustering paradigms. The evaluation metrics included a complete set designed specifically for the data mining tasks, which quantified model performance.

The designed experimental procedure enables a methodical analysis of pre-processing methods on mining results while showing their effectiveness across multiple datasets and analytical applications. Data preprocessing techniques solve missing values together with noise, inconsistency, and redundancy problems, improving the accuracy, efficiency, and interpretability of data mining models [2-4]. Equipping impractical datasets for data mining requires data preprocessing, reflecting their input parameters to diverse data mining algorithms. Researchers and business activities depend on advanced data analysis systems because modern data volumes keep expanding [3].

The data preprocessing process includes four main activities: data cleaning, transformation, and reduction before data integration [6]. The data cleaning process includes three stages: handling missing values, outlier detection, and inconsistency correction in the data. The methods used to handle data gaps in datasets include mean substitution and median substitution imputation with advanced techniques utilizing k-nearest neighbors imputation. Detecting outliers through z-score analysis, box plot analysis, and clustering-based methods enables users to identify and discard or reduce the impact of deviation points. The data conversion process requires scaling alongside normalization and feature engineering alterations to produce optimal input data for data mining algorithms. Data reduction methods consist of dimensionality reduction and feature selection, which decrease variable or feature count while maintaining critical knowledge in the database.

The integration process incorporates multiple data resources to generate a single unified dataset while it handles conflicts and superfluous information. Data preprocessing substantially affects data mining results because it directly controls their outcome quality and reliability [2]. Knowledge discovery becomes very difficult when the presented data contains excessive amounts of redundant material, unrelated data noise, and unreliable information [4]. The standard of data preprocessing determines how precise and dependable models become, so poorly processed data might generate wrong findings.

Data preprocessing success rates are determined by how datasets have been structured and the desired outcomes for data mining. The performance of classification algorithms gets evaluated in extended benchmark exams, which use standardized datasets for testing predictive accuracy and computational performance [2]. Although no single preprocessing algorithm works best for every dataset, multiple techniques yield improved success in data mining projects [4].

Research shows that data preprocessing is essential in discovering knowledge from data sources [3]. The data compilation process stands as a controlled activity, and data typically contains inconsistencies with errors and out-of-range values, leading to impossible data combinations, missing or inappropriate values, and invalid data that bars data mining commencement. According to Alexandropoulos et al. (2019), machine learning processes require extensive time to execute their data preparation protocols. Various preprocessing algorithms cannot deliver reliable and effective performance for every dataset because this becomes impossible.

4. Results and Discussion

Data preprocessing profoundly affects the operational performance of data mining algorithms, which were experimentally tested among multiple datasets and analytical functions. The model accuracy and robustness substantially improved when missing values and outliers were addressed through proper imputation followed by removal procedures. Processing methods that normalize and scale data features managed to overcome fluctuations in data magnitudes, resulting in better model behavior.

Developing new features or transforming data characteristics through Feature Engineering produced significant advantages by enhancing prediction capabilities and interpretability. Model complexity became simpler using PCA or feature selection methods, which improved computational efficiency by maintaining critical data information.

The observed impact consists of these particular examples:

Combining proper data preprocessing methodologies that manage unbalanced data followed by feature selection resulted in substantial improvements in classification algorithms, producing better customer churn and fraud detection predictions.

Analysts applied three techniques to data, which improved regression models for house price estimation and stock market direction forecasting.

The quality of customer segmentation and anomaly detection clusters advanced due to the implementation of normalization methods combined with dimensionality reduction techniques.

The findings validate the essential position that data preprocessing holds in achieving optimal, reliable outcomes from data mining activities.

Researchers uncovered that preprocessing data is critical in boosting both performance and reliability levels of mining data results. Data preprocessing is a complex requirement for data mining in database knowledge discovery procedures [2]. Data preprocessing methods handle the problems of data gaps and extreme values, hierarchy variations, and undesirable characteristics, resulting in higher analytic data quality. A domain selection directly shapes which dataset becomes the focus. The essential element is choosing appropriate target data, while the selected data should be taken from the raw dataset to improve confidence [6].

The data preprocessing process allows algorithms to find valuable patterns, which leads to an enhanced understanding of the data. The main focus of DPP research is developing algorithms that solve specific tasks within the field. Data preprocessing makes the data possible for various data mining algorithms by conforming to input requirements [3]. Decision-making about preprocessing techniques becomes crucial because data characteristics and mining task objectives require individual analysis [6]. Excessive and inadequate preprocessing methods can cause model performance issues and discard important data or introduce biases in the analysis [4].

Data preprocessing techniques include Data Cleaning followed by Data Optimization, Data Transformation, and Data Integration with Data Conversion as the last step. The applied data preprocessing methods improved model accuracy and robustness and enhanced interpretability, proving the practical value of deep data preprocessing approaches [6]. Different preprocessing methods produce variable results based on data characteristics and mining goals from specific tasks.

The required process of converting unusable data into usable form through data transformation follows each data mining algorithm's input specifications. The data preprocessing techniques create data sets of excellent quality that support data analysis requirements, improve data quality, strengthen model performance, and ensure reliable data mining results [3,4,6].

5. Conclusion

Data preprocessing stands as an essential component that guarantees the achievement of data mining goals. Through data preprocessing techniques that handle quality problems, format data correctly, and reduce dimensions, researchers can boost the accuracy by lowering run-time and improving result interpretations of models [3, 5,6]. The analysis evaluates data preprocessing's role in enhancing analytical results to provide better quality and more reliable outcomes at higher efficiency levels.

The representation and dataset quality emerge as the two essential operative matters. The experimental outcome demonstrates the need to use data preprocessing methods that match the dataset's characteristics and the data mining purpose. Data preprocessing allows algorithms to discover accurate predictions through its systematic approach to fix data quality issues alongside representation transformation and dimension reduction.

References

[1] Kasula, V. K. (2022). Empowering Finance: Cloud Computing Innovations in the Banking Sector. International Journal of Advanced Research in Science Communication and Technology, 2(1): 877-881, http://dx.doi.org/10.48175/IJARSCT-12467I

[2] Crone, S. F., Lessmann, S., & Stahlbock, R. (2005). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. In European Journal of Operational Research (Vol. 173, Issue 3, p. 781). Elsevier BV. https://doi.org/10.1016/j.ejor.2005.07.023

[3] García, S., Luengo, J., & Herrera, F. (2015). Tutorial on practical tips for the most influential data preprocessing algorithms in data mining. In Knowledge-Based Systems (Vol. 98, p. 1). Elsevier BV. https://doi.org/10.1016/j.knosys.2015.12.006

[4] Alexandropoulos, S.-A. N., Kotsiantis, S., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. In The Knowledge Engineering Review (Vol. 34). Cambridge University Press. https://doi.org/10.1017/s026988891800036x

[5] Famili, A. F., Shen, W.-M., Weber, R. W., & Simoudis, E. (1997). Data Preprocessing and Intelligent Data Analysis. In Intelligent Data Analysis (Vol. 1, Issue 1, p. 3). IOS Press. https://doi.org/10.3233/ida-1997-1102

[6] Joshi, A. P., & Patel, B. V. (2021). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. In Oriental journal of computer science and technology (Vol. 13, Issue 203, p. 78). https://doi.org/10.13005/ojcst13.0203.03

[7] Li, C. (2019). Preprocessing Methods and Pipelines of Data Mining: An Overview. In arXiv (Cornell University). Cornell University. https://doi.org/10.48550/arxiv.1906.08510

[8] Cases, P. U., & Tolulope, A. I. (2022). Data Science and Analytics for SMEs.

[9] Nereu, J. F. C. (2017). Open Source Platforms for Big Data Analytics (Master's thesis, Instituto Politecnico do Porto (Portugal)).