A Comprehensive Survey and Framework for Handling Noisy and Incomplete Data in Machine Learning

V.Balu
Assistant Professor
Dept of CSE
SCSVMV (Deemed to be University)
Kanchipuram

Dr.C.K.Gomathy
Assistant Professor
Dept of CSE
SCSVMV(Deemed to be University)
Kanchipuram

Abstract
In today's data-driven world, real-world data is often noisy, incomplete, inconsistent, and heterogeneous in nature ranging from structured numerical records to unstructured media formats such as text, images, audio, and video. These imperfections in data quality pose significant challenges for data mining tasks and predictive modeling. This study presents a comprehensive survey of techniques used to handle missing and noisy data, highlighting the advantages and limitations of methods such as imputation (mean, median, and regression), outlier detection, and noise filtering. Through literature review and empirical findings from various machine learning models—such as SVM, KNN, Random Forest, and ensemble methods. This paper identifies the effectiveness of hybrid approaches in improving data quality and model accuracy. The proposed work suggests a systematic architecture that includes preprocessing, imputation, classification, and evaluation phases, applied to real-world datasets. This framework aims to enhance data reliability, reduce false predictions, and increase the efficiency of data mining applications.
Keywords: Data mining, Inconsistent data, scalable, Data cleaning, decision making, data quality.

I. Introduction

In the current digital era, vast volumes of data are generated from diverse sources such as databases, data warehouses, sensor networks, and online platforms. However, this data is often noisy, incomplete, inconsistent, and heterogeneous, making it challenging to extract meaningful insights. The data can exist in multiple forms, including numerical values, categorical variables, time series, spatial and temporal data, audio, video, images, and natural language text [3][9]. These varied formats, combined with human error, technical malfunctions, and privacy concerns, often result in low-quality datasets that compromise the reliability of downstream data mining and analysis tasks.

Noisy data, which includes errors, outliers, or inconsistencies, acts as meaningless data and reduces the quality of analysis. "This type of noise can originate from errors during data entry, malfunctioning hardware, or irregularities in the data gathering process." Removing noisy samples before training machine learning models can significantly enhance data quality [3][9]. Similarly, missing data occurs when values are not recorded due to issues like system crashes, non-responses in surveys, data corruption, or privacy-preserving mechanisms [9]. Missing values appear in both numerical and categorical columns, posing a risk to the accuracy of prediction models and the overall integrity of data-driven decisions.

Data mining plays a critical role in uncovering useful patterns and knowledge from large datasets. "Nevertheless, its performance is hindered by the existence of noise and incomplete data." Various studies have applied machine learning models to handle such issues. For instance, Support Vector Machine (SVM) regression has been employed for imputing missing values, achieving high classification accuracy when applied to the PIMA Indian diabetes dataset [1]. Hybrid approaches combining techniques like K-Nearest Neighbors (KNN), clustering, and ensemble learning have also been proposed to enhance both data imputation and classification performance [2][6][7][11].

This research presents a comprehensive survey of the approaches used to address noisy and incomplete data, discussing their strengths and limitations. It highlights traditional and advanced methods such as mean/median imputation, regression models, and machine learning algorithms for noise filtering and data prediction. Ultimately, the goal is to improve data quality, algorithm performance, and the accuracy of data mining results, while also considering scalability, efficiency, and data privacy challenges [4][7][9].

II. Literature Review

Numerous studies have explored methods to handle noisy and incomplete data using machine learning and data preprocessing techniques. A variety of algorithms and hybrid approaches have been proposed to improve classification performance, missing data imputation, and noise filtering across different application domains and datasets.

In [1], the authors compared the performance of five machine learning models—Naive Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Random Forest (RF), and Linear Regression (LR)—on the PIMA Indian Diabetes dataset. "Support Vector Machine (SVM) regression was employed to fill in missing values, followed by a two-stage classification approach to minimize incorrect classifications." The study reported that the SVM classifier achieved the highest accuracy of 94.89%, RF showed the highest precision (98.80%), and NB recorded the best F1-Score (95.59%). SVM also demonstrated the highest recall at 85.48%, highlighting its effectiveness in handling missing data and classifying diabetic outcomes.

Several studies [2][6][11] utilized the KNN method for missing value imputation and found it to outperform traditional statistical techniques such as mean, median, and mode. These approaches further incorporated k-means clustering for noise detection, followed by the application of various classification filtering algorithms to enhance data quality. The use of hybrid models was found to be a more robust solution for both missing value imputation and classification tasks. Additionally, the computational cost of these methods was evaluated and monitored to ensure scalability and practical applicability.

In [4], the author proposed a Noise-Aware Multiple Imputation (NPMI) algorithm based on a random sampling consistency strategy. This method was tested on real-world datasets, including weather and sensor data, and the results showed significant improvements in both the accuracy and efficiency of missing data handling.

The study in [7] aimed to enhance healthcare predictions by combining KNN-based imputation with a TriEnsemble model comprising Random Forest, Extra Trees Classifier, and XGBoost. This ensemble approach achieved an accuracy of 94.7% for diabetes prediction and suggested that future work could involve integrating deep learning models for further performance gains.

In [9], the authors evaluated supervised learning methods such as KNN, SVM, and Classification and Regression Trees (CART) for handling noisy and incomplete data. Among these, CART performed best in terms of classification accuracy. The paper also proposed the application of ensemble learning techniques in future studies to construct more resilient imputation models.

Various datasets have been employed in these studies, covering domains such as healthcare (e.g., PIMA Indian Diabetes, Cleveland Heart Disease, Breast Cancer), environment (e.g., Weather, Air Quality, Ground Water, Pollution Data), and user-generated content (e.g., TripAdvisor Reviews, Grades Data). These datasets were collected from reputable sources including the UCI Machine Learning Repository, Kaggle, TripAdvisor.com, PIMA India Data Repository, BMS Data, EUI Survey Data, and the ASHRAE database. The diversity in datasets demonstrates the wide applicability and relevance of noise reduction and missing data imputation strategies in real-world scenarios.

2.1 Problem Statement

In contemporary data-driven environments, the presence of incomplete and noisy data poses a significant challenge to effective data analysis and decision-making. Incompleteness may arise from various sources, including data collection errors, integration mismatches, limitations in data preprocessing, privacy-preserving mechanisms, or the natural evolution of datasets over time. Noisy data, on the other hand, often contains inaccuracies, outliers, and inconsistencies introduced through human error, sensor faults, or technical anomalies. These issues not only reduce the quality of the data but also undermine the effectiveness of downstream tasks such as classification, prediction, and pattern recognition.

The implications of poor data quality are profound: noisy and incomplete datasets can lead to misleading analytical outcomes, reduced model accuracy, and decreased efficiency and scalability of data mining algorithms. Moreover, certain data cleaning operations—such as outlier removal or duplicate filtering—may inadvertently eliminate valuable information, further degrading the dataset's integrity.

This research investigates the problem of handling noisy and incomplete data in heterogeneous and large-scale datasets. It presents a comprehensive survey of state-of-the-art imputation techniques and noise reduction strategies used in machine learning and data mining. The study critically evaluates the advantages and limitations of each approach in terms of accuracy, computational cost, and practical applicability. The key contribution of this research lies in identifying optimal methodologies for improving data reliability and mining performance while addressing constraints related to data privacy, algorithmic efficiency, and domain-specific requirements.

III. Missing data

Some variables or observations have missing data due to unavailability or lack of recording. When a dataset contains missing values, it is referred to as an incomplete dataset[9]. i.e Few datasets may have incomplete attribute values, lack key features, or include only aggregated information. Lost data may appear in columns containing either numerical or categorical values. Missing data can be caused by various factors, such as Technical Problem, Lack of data Observation, User Privacy Issues, Human Error, Incomplete surveys, Non-response, Data loss, or Data corruption, program errors etc. Missing data can lead to gaps in the data mining process, potentially compromising the validity and reliability of the results. It will affect the accuracy of the predictions. The below diagram figure 1 is represented missing values in numerical columns

Fig 1. Missing Values in Numerical Columns

Types of missing data

- Data missing at random(MAR)
- Data missing completely at random(MCAR)
- Non ignoring missing data
- Outliers treated as missing data

Missing Values in Categorical Columns

Missing values in categorical columns is generally simpler compared to numerical values. A common method involves replacing missing entries with a fixed value or the most frequent category, which is particularly effective for smaller datasets. For instance, if an 'Education' column contains values like 'High School' and 'College,' and the majority of entries are 'College,' it's reasonable to fill in the missing values with 'College.'"

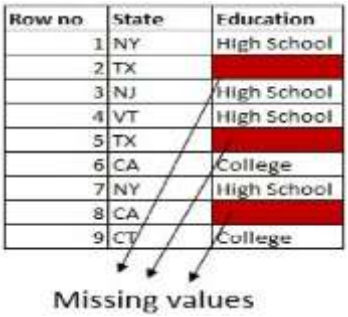

Fig 2. Missing values in Categorical Columns

"We can improve the handling of missing values by utilizing information from other columns. For instance, if most individuals from Texas have 'High School' listed as their education level, it's reasonable to fill in missing values for Texas residents with 'High School.' "An alternative approach is to use a classification model to estimate the missing 'Education' values by analyzing patterns in the other features of the dataset." However, a widely used strategy is to treat the missing category as a separate label, such as 'Unknown.' Various techniques—like using the mean, median, or regression—are also commonly applied to handle missing data.

3.1 Missing Data Approaches

Figure 3 illustrates the various approaches available for handling missing values.

Fig 3. Missing value Treatment

Data imputation methods such ad Drop, Mean imputation, median imputation, or regression imputation to fill in or replace any missing data with estimated or predicted values. Let us consider above example fig 1 and fill the missing values using with Mean imputation method represented in fig 4.

Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX	65150	Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX	65150	Mid
6	CA	76600	High
7	NY	85000	High
8	CA	65150	Entry
9	CT	45000	Entry

Replaced with the mean salary

Fig 4. Mean Computation Method

Mean(x)=Total value/No of available records ---(1)

One basic method for handling missing values is to use substitution strategies such as:

- Using the mean or median of the column, this is particularly helpful when the dataset is small.
- Filling in missing values by leveraging related information from other columns.

For an example, in an employee dataset where the Salary column has missing values in three rows, we might begin by filling in those gaps using the overall average salary. However, with guidance from a domain expert, we can refine this approach further. Since average salaries often vary by location, we could compute the average salary for a specific state—like Texas—and use that figure to fill in missing values for individuals from that state.

We can improve this further by considering additional attributes, such as Years of Experience. For instance, if an entry-level employee from Texas has a missing salary, we could use the average entry-level salary specific to Texas. The same logic applies to mid-level and senior-level roles.

There are also edge cases to consider. For example, if both Salary and Years of Experience are missing for a row, one practical solution is to fill the missing salary with the average for that state—such as the average salary in Texas.

Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX	45000	Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX	45000	Mid
6	CA	76600	High
7	NY	85000	High
8	CA	55000	Entry
9	CT	45000	Entry

Replaced with mean salary in TX

Replaced with mean salary in CA

Replaced with mean entry level salary in TX

Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX	35000	Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX	48000	Mid
6	CA	76600	High
7	NY	85000	High
8	CA	43000	Entry
9	CT	45000	Entry

Replaced with mean Mid level salary in TX

Replaced with mean Entry level salary in CA

3.2. Predicting Missing Values Using a Model

Another approach involves using a predictive model, such as linear regression, to estimate the missing values. The goal here is to estimate the missing Salary entries by utilizing the information from other features in the dataset. If some input features also contain missing values, need to manage those cases—either by selecting only features without missing data or by using rows that are fully complete to train the model.

3.3 Handle Noisy data

- Data cleaning Methods (data validation, normalization, transformation, or correction can be employed to remove or fix any errors or inconsistencies and improve the accuracy.

- Outlier detection Methods like z-score, inter quartile range, or clustering can be used to identify and separate any outliers or extreme values.

- Noise filtering methods like smoothing, binning, or regression can be applied to reduce or eliminate any noise

- Factors to be considered Type (Categorical / Numerical), source, pattern, and impact of the noisy and missing data

IV Proposed Framework

The Architecture of the proposed work is represented in Fig 7. In this proposed work, missing values can be handled by applying deletions or imputations method

Step1: Ignore observations of missing values if dealing with large datasets and less number of records has missing values.
Step2: Replace with the most frequent values with the mean /median or KNN method
Step3: Split the dataset as 80 %training dataset&20%testing dataset
Step4: Apply classification algorithm/ regression /Decision tree algorithm
Step5: Compare the performance of algorithm with hybrid approach
Step6: Predict the result

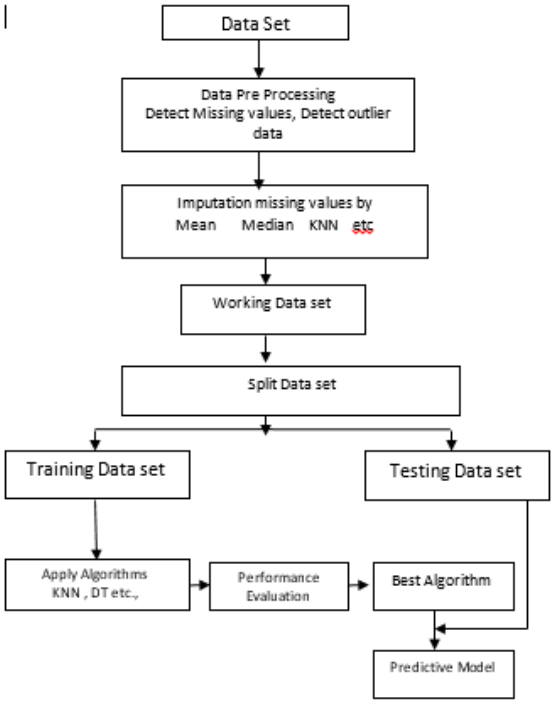

Fig 7: System Architecture

XGBOOST algorithm is a supervised regression model is used for predicting accuracy of the system. Decision Tree algorithm is used for predicting Target value. Gaussian Naïve Bayes algorithm output of an event is predicted using unconditional probabilities. Perform two level classification processes to reduce the number of false classification. Compare the performance of the different machine learning models and to be choose the best algorithm for effective system

Conclusion

Handling noisy and incomplete data is crucial for ensuring the quality and reliability of data mining outcomes. This research emphasized the impact of data imperfections on predictive performance and surveyed a range of imputation and noise-handling techniques from existing literature. It was observed that machine learning-based imputation, combined with hybrid ensemble classifiers, offers improved accuracy and resilience to data quality issues. The proposed framework, which systematically addresses missing data through deletion, imputation, and predictive modeling, demonstrates a practical

pathway to enhance real-world data mining applications. Future work will focus on integrating deep learning-based imputers and scalable algorithms to handle high-dimensional, real-time, and privacy-sensitive datasets.

References

1. Ashokkumar Palanivinayagam and Robertas Damaševiˇcius, Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods Information on Open Access Journal from MDPI **(2023)**

2. Niyaz Sharifyanov, Viktoriya Latypova ”A Method of Filling Missing Values in Data using Data Mining” IX International Conference on Information Technology and Nanotechnology (ITNT) IEEE Explore **(2023)**

3. Miss VaidyaVijayshri Dattatray, Mr.Nibe Abhishek Annasaheb A Review on Data Mining and Data Preprocessing Techniques in Data Mining IJARIIE **(2023)**

4. Fangfang Li*, Hui Sun, Yu Gu and Ge Y “A Noise-Aware Multiple Imputation Algorithm for Missing Data” Journal of Mathematics **(2022)**

5. NapsuKarmitsa , SonaTaheri, AdilBagirov, and Pauliina Makinen Missing Value Imputation via Cluster wise Linear Regression IEEE Transactionsonknowledgeanddataengineering,vol.34,no.4,april **(2022)**

6. Monalisa Jena Andsatchidanandadehuri An Integrated Novel Framework for Coping Missing Values Imputation and Classification IEEE Access **(2022)**

7. Khaled Alnowaiser “Improving Healthcare prediction of diabetic using KNN imputed features and TriEnsemble Model” IEEE Access(2024)

8. Tlamelo Emmanue* , ThabisoMaupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago and Oteng Tabona A survey on missing data in machine learning Joural of Bigdata Springer **(2021)**

9. Ya-Han Hu, Chih-Fong Tsai “An investigation of solutions for handling incomplete online review datasets with missing values” Journal of Experimental & Theoretical Artificial Intelligence **(2021)**

10. Cheng Fan , Meiling Chen , Xinghua Wang et al “A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data “Front. Energy Res Sustainable Energy Systems (**2021)**

11. Zahra Nematzadeh ,· Roliana Ibrahim et al.” A hybrid model for class noise detection using k-means and classification filtering algorithms” SN Applied Sciences **(2020)**

12. L. Sunithaa*, M.Bal Rajua and B.Sunil Srinivasa “A comparative study between noisy data and outlier data in data mining”, International journal of current engineering and Technology Vol.3, No.2 (June 2013)